2017

Validity and Reliability of Baseline Testing in a Standardized Environment

Kathryn L. Higgins
University of Nebraska–Lincoln, drkatehiggins@gmail.com

Todd Caze
University of Nebraska–Lincoln, tcaze2@unl.edu

Arthur C. Maerlender
University of Nebraska-Lincoln, amaerlender2@unl.edu

Published in *Archives of Clinical Neuropsychology,* 2017. doi: 10.1093/arclin/acx071

Accepted 25 July 2017.

Validity and Reliability of Baseline Testing in a Standardized Environment

Kathryn L. Higgins, Todd Caze, and Arthur Maerlender

Center for Brain, Biology & Behavior, University of Nebraska–Lincoln

Corresponding author — K. Higgins, Center for Brain, Biology & Behavior, University of Nebraska–Lincoln, C64, East Stadium, PO Box 880156, Lincoln, NE 68588, USA; email drkatehiggins@gmail.com

Abstract

Objective — The Immediate Postconcussion Assessment and Cognitive Testing (ImPACT) is a computerized neuropsychological test battery commonly used to determine cognitive recovery from concussion based on comparing post-injury scores to baseline scores. This model is based on the premise that ImPACT baseline test scores are a valid and reliable measure of optimal cognitive function at baseline. Growing evidence suggests that this premise may not be accurate and a large contributor to invalid and unreliable baseline test scores may be the protocol and environment in which baseline tests are administered. This study examined the effects of a standardized environment and administration protocol on the reliability and performance validity of athletes' baseline test scores on ImPACT by comparing scores obtained in two different group-testing settings.

Method — Three hundred-sixty one Division 1 cohort-matched collegiate athletes' baseline data were assessed using a variety of indicators of potential performance invalidity; internal reliability was also examined.

Results — Thirty-one to thirty-nine percent of the baseline cases had at least one indicator of low performance validity, but there were no significant differences in validity indicators based on environment in which the testing was conducted. Internal consistency reliability scores were in the acceptable to good range, with no significant differences between administration conditions.

Conclusions — These results suggest that athletes may be reliably performing at levels lower than their best effort would produce.

Keywords: Neuropsychological testing, Collegiate athletes, Baseline, Concussion, Concussion assessment

Introduction

As sport-related concussion injuries gain notoriety and management techniques work to keep up, clinical practices have been put into place that may or may not have strong evidence-based support. Baseline testing began in the 1980s and now is common practice in most sports concussion management programs. The individual's effort in testing has long been recognized as an important moderator in neuropsychological testing outcomes, and within the last 10 years has been studied in the context of baseline testing. Both environmental factors (e.g., variable and distraction-filled testing rooms) and individual (e.g., lack of motivation and fatigue) have been shown to contribute to poor testing outcomes in baseline testing (McCrory, Makdissi, Davis, & Collie, 2005). These threats to consistent and optimal performance raise important questions about the rates of potential invalidity and internal reliability of baseline testing scores.

Baseline Testing in Computerized Neurocognitive Testing

Baseline testing provides an individualized reference point to guide return to play decisions based on cognitive recovery (Barth et al., 1989), in contrast to norm-based test interpretation, which utilizes normative data to interpret an individual's test data. While the baseline model appears very theoretically sound, there is limited empirical support for it (Echemendia et al., 2012), limited evidence that it modifies risk (Randolph,

2011), can be very time and labor intensive (Iverson & Schatz, 2015), and "may be useful, but is not necessary for interpreting post-injury scores" (McCrory et al., 2017, p. 3). For example, Echemendia et al. determined using normative cut-offs demonstrated sensitivity of 80%–86% and specificity of 95%–97% for identifying clinically meaningful cognitive change after concussion when compared to a baseline model. In contrast, Schatz and Robertshaw (2014) found that normative scores were significant more likely to misclassify above average athletes as uninjured when compared to baseline/post-injury comparisons.

Return to play guidelines strongly recommend that athletes not be returned to play until they are completely asymptomatic and fully recovered cognitively (Broglio et al., 2014; Harmon et al., 2013), and baseline testing can help to operationalize the meaning of "fully recovered" for individual athletes (Piland et al., 2010). But, as pointed out by Erdal (2012), the utility of the comparison between post-injury and baseline test data in return-to-play decisions is based upon the integrity of the baseline data.

ImPACT Baseline Validity

The importance of assessing effort is frequently discussed in general neuropsychological testing, and is gaining increased attention in baseline testing for concussion management; research has demonstrated that athletes can intentionally or unintentionally alter their baseline scores, based on the amount of effort that they put forth at the time of baseline testing (Bailey et al., 2006).

ImPACT includes several built-in indicators of potentially invalid baseline data. A systematic review by Gaudet & Weyandt (2016), found invalidity rates (based on ImPACT invalidity indicators) in normal baseline samples ranging from 2.7% to 27.9%, with a weighted prevalence rate across the 12 studies of 6.1% (Gaudet & Weyandt, 2016). Performance validity research using simulators has demonstrated that the invalidity indicators built into ImPACT do not identify all those who provide suboptimal effort (colloquially termed "sandbagging"). Research utilizing ImPACT-savvy and ImPACT-naïve participants of a variety of ages has found between 11% and 35% of those asked to provide suboptimal effort were able to do so without being identified by ImPACT's invalidity indicators (Erdal, 2012; Schatz & Glatts, 2013; Higgins, 2015).

In addition to ImPACT invalidity indicators, Schatz & Glatts (2013) found that a Word Memory Correct Distractors score <22 and a Design Memory Correct Distractors score <16 had the high utility and identified 95% and 90% of naïve "sandbaggers" and 100% and 95% of coached "sandbaggers", respectively. These two variables were more effective than an established performance validity test (Medical Symptom Validity Test), which identified 80% and 90% of naïve and coached "sandbaggers", respectively. Higgins, Denney, & Maerlender (2017) also determined that a Logistic Regression Equation ≥0.23 utilizing several ImPACT subtest scores demonstrated 100% sensitivity and 90% specificity to suboptimal effort.

ImPACT Baseline Reliability

Schatz's (2010) study of collegiate athletes' ImPACT baselines (taken at a 2-year interval) found 2 year test–retest reliability to be higher and in the "good" range for speed composites (Processing Speed ICC = .74; Reaction Time ICC = .68) and Visual Memory composite (ICC = .65), than for Verbal Memory composite (ICC = .46). Other studies have questioned the acceptability of ImPACT test–retest reliability (citing .70 as being the normally applied cutoff for acceptable reliability) and noted the differences in research testing intervals (2 weeks to 1 year) and clinical testing intervals (up to 2 years in high school students and 4 years in college students) (Mayers & Redick, 2012).

Standardized Administration and Environment in Testing

ImPACT does not come with standardized instructions to be given to athletes before test administration, which may contribute to variability and increased error (Moser, Schatz, Neidzwski, & Ott, 2011). ImPACT can be administered in a group or individualized setting, and while group administration may be more appealing due to fewer time and personnel requirements, it can also be detrimental to the validity of test data (Moser et al., 2011).

In contrast, Vaughan, Gerst, Sady, Newman, & Gioia (2014) found that when test administrators were trained in neuropsychological testing and the test administration was standardized, there was no difference between scores obtained in group versus individualized testing for a group of 5–18 years old participants.

Factors specific to individual athletes (e.g., learning disabilities, fatigue, stress, etc.) can have a significant effect on cognitive scores and can be more difficult to track and potentially compensate for in group testing situations, particularly when administration procedure is not scripted or standardized. It is also recommended

that baseline tests not be taken at a time when the athlete is sick, tired, very stressed, or experiencing a physical (e.g., orthopedic) injury (Moser, Schatz, & Lichtenstein, 2013; Piland et al., 2010).

The aims of this study were to examine how a standardized environment and administration protocol affect the reliability and performance validity of athletes' baseline test scores on ImPACT in two different group-testing setting. The specific hypotheses were that when compared to an unstandardized procedure, a standardized environment and administration protocol would result in fewer baseline tests with indicators of potential invalidity, and higher internal consistency (reliability).

Method

Participants

Participants completed baseline testing as a part of the university athletics' concussion management program.

From a large dataset of 1077 ImPACT records collected from August 2, 2012 to January 22, 2016, a set of 778 cases were identified (299 were post-injury tests). Rule-outs from that set were the presence of a learning disability (n = 9), ADHD (n = 34), a second test by the same athlete (n = 27), and non-athlete cases (n = 53), leaving 655 cases.

Informed consent was waived by the Institutional Review Board as all data was de-identified by an honest broker before analysis and was not considered to be human subjects' data.

Materials

All ImPACT tests were in Version 2.1 (ImPACT Application, Inc., 2012). The current version of the ImPACT generates five composite scores: Verbal Memory Composite, Visual Memory Composite, Reaction Time Composite, Impulse Control Composite, and Total Symptom Composite Score. Psychometric properties of ImPACT are reviewed above.

Procedures and Analyses

In 2014, a standardized protocol for baseline administration was implemented. Prior to that, baseline testing was conducted by athletic trainers and graduate assistants with no consistent procedure or protocol. Those tested before standardization are referred to as Group 1, and those after standardization as Group 2.

Testing for Group 1 was completed in a computer lab in the Athletics facility. Each athletic trainer was responsible for baseline testing the athletes on their teams, which resulted in both group and individual baseline test with no specified or consistent administration protocol. No other information about procedures was available.

For Group 2, baselines were administered using a standard script and test administration procedure. Testing groups were limited to 15 athletes with an athlete-to-proctor ratio at or below 5:1. Testing for Group 2 was completed in a university computer lab with each computer located in an individual carrel. Athletes were gathered in a waiting area and given the standardized script. They were given the opportunity to ask questions and were then asked to turn off their cell phones.

Because there were differences between groups in the number of contact sport athletes, and sex, a cohort matching strategy was utilized based on sport (contact vs. non-contact), sex, age (within 1 year), and total symptom score (within three points), resulting in Group 1 n = 178 and Group 2 n = 183. Total symptom score was included as a matching criteria, as preliminary analysis of the sample (before cleaning and matching) demonstrated a small, but statistically significant difference between the two groups on total symptom score.

Several performance validity indicators were employed to assess the quality of the test results: that provided by ImPACT ("baseline++" notation), the two indicators found to have greatest utility by Schatz and Glatts (2013), and Higgins and colleagues (2017) logistic regression equation. Each specific indicator was coded as either 0 (test results were valid) or 1 (test results were of questionable validity). The number of indicators that were met for each case was tabulated. Group means on cognitive composites and total symptom score were then compared using a t-test or Mann–Whitney U, depending on the distribution of the data. Level of significance was adjusted using Bonferroni correction to control for Type 1 error.

While test–retest reliability is the reliability assessment commonly used in ImPACT research, this study's focus on only baseline testing made internal consistency reliability the more appropriate metric. In order to be able to assess internal reliability within constructs, all subtests scores were converted to z-scores. Scores for which a lower number indicates a better performance (such as reaction time scores) were reverse-scored

to normalize directionality (Pallant, 2005). Summary scores based on subtest score combinations were not included (i.e., Word Memory Learning Percent Correct). The scores included in the reliability analysis are shown in Table 1.

To assess internal consistency within each group, Cronbach's alpha was calculated for all standardized subtest scores for Group 1 and Group 2. Cronbach's alpha for memory and speed subtest scores were calculated using the subtest variables notated with a 1 (memory) or 2 (speed) in Table 1. A Fisher's *z*-test was then used to compare the reliability scores between the two groups.

Results

Participants' average age was 19.2 (1.4) years in Group 1 and 18.9 (1.5) years for Group 2. About half of the sample were contact athletes (Group 1 = 54.5% and Group 2 = 47.0%) and about two-thirds were male (Group 1 = 69.7% and Group 2 = 61.2%). No significant differences were found between Group 1 and Group 2 on any ImPACT composite, hours of sleep, or number of previous concussions.

Performance Validity Analysis

The numbers of cases identified by each of the invalidity indicators are listed in Table 2. There were no significant differences between groups in the number of cases identified by the potential indicators of invalidity (Group 1 = 38.8% and Group 2 = 31.1%; χ^2 (1, N = 361) = 2.30, p = .13). Of those cases in both groups identified as potentially invalid, a majority were identified by 1 indicator only (n = 126; 34.9%). Still, 21% of Group 1 and 17% of Group 2 had 2+ indicators of potential invalidity.

Reliability Analysis

No significant differences were found between Group 1 and Group 2 based on the internal consistency within all of the subtest scores; internal reliability was in the "good" range for both groups (see Table 3). There was also no difference between groups when just the 17 memory subtest scores were examined, and reliability

Table 1. Subtest used in reliability analysis

Subtests	
1	Word Memory Hits (Immediate)
1	Word Memory Correct Distracters (Immediate)
1	Word Memory Hits (Delay)
1	Word Memory Correct Distracters (Delay)
1	Design Memory Hits (Immediate)
1	Design Memory Correct Distracters (Immediate)
1	Design Memory Hits (Delay)
1	Design Memory Correct Distracters (Delay)
1	X's and O's Total Correct (Memory)
2	X's and O's Total Correct (Interference)
	Symbol Match Total Correct (Visible)
1	Symbol Match Total Correct (Hidden)
	Color Match Total Correct
1	Three Letters Total Sequence Correct
1	Three Letters Total Letters Correct
2	Three Letters Average Counted
2	Three Letters Average Counted Correctly
Reverse-score Subtest Scores	
2	X's and O's Average Correct Reaction Time (Interference)
	X's and O's Total Incorrect (Interference)
	X's and O's Average Incorrect Reaction Time (Interference)
2	Symbol Match Average Correct Reaction Time (Visible)
2	Symbol Match Average Correct Reaction Time (Hidden)
2	Color Match Average Correct Reaction Time
	Color Match Total Commissions
	Color Match Average Commissions Reaction Time
2	Three Letters Average Time to First Click

1 = memory factor, 2 = speed factor.

Table 2. Number of cases (%) identified by each performance validity indicator by group and significance

Indicator	Group 1	Group 2	Chi square
Baseline ++	5 (2.8%)	4 (2.2%)	Not significant
Word Memory Correct Distractors <22 (Schatz and Glatts, 2013)	32 (18.0%)	20 (10.9%)	Not significant
Design Memory Correct Distractors <16 Schatz and Glatts, 2013)	44 (24.7%)	41 (22.4%)	Not significant
Logistic Regression Equation ≥0.23 (Higgins et al., 2017)	41 (23.0%)	37 (20.0%)	Not significant
Total cases identified[a]	69 (38.8%)	57 (31.1%)	Not significant

a. Total cases will not be the sum of all cases identified, as some cases had more than one indicator.

Table 3. Reliability analysis

Scores included	Group 1 Cronbach's α	Group 2 Cronbach's α
All subtest scores (26)	0.82	0.84
Memory subtest scores (17)	0.80	0.83
Speed subtest scores (9)	0.61	0.61

was in the "good" range for this subgroup of subtest scores. Finally, there was also no difference between Group 1 and Group 2 on the nine speed subtest scores, but reliability for this subgroup of scores was in the "questionable" range.

Discussion

The aims for this study were to evaluate how environment and administration protocol affected the performance validity and internal consistency (reliability) of athletes' baseline test scores on ImPACT when obtained in group testing settings. Moser and colleagues (2011) demonstrated that athletes tested in a group setting demonstrate poorer cognitive composite scores and higher rates of invalid baselines than athletes tested individually. Vaughan and colleagues (2014) extended that research by demonstrating that athlete tested in a well-controlled group environment with a standardized procedure demonstrated no differences on cognitive composites or invalidity rates than those tested individually. Here we compared two large samples that differed by the administration protocol and environmental control in which baseline testing was conducted.

Our hypotheses that a standardized administration environment and protocol would decrease the number of cases with suspected performance invalidity, while improving reliability as measured by internal consistency were both unsupported. In this matched cohort sample of Division 1 athletes without premorbid developmental concerns, the performance validity data suggests that many athletes may be performing at a level that is lower than their best effort would produce. Using a variety of indicators, we found that 31%–39% of these baseline cases had at least one indicator of low performance validity and 17%– 21% had two or more indicators of invalidity. Literature on normal score variability has clearly demonstrated that with multiple data points, having at least one low score is the rule, rather than the exception (Binder, Iverson, & Brooks, 2009). With that in mind, the invalidity indicators used in this analysis are not random low scores, but rather low scores that are shown to predict poor effort. Also, Larrabee (2012, 2016) warns that utilizing more than one effort (validity) measure can increase the risk of misidentifying someone providing sufficient effort. These factors may have a role in the number of potential cases of invalidity identified. With those important caveats, the findings from this study demonstrate that both groups demonstrated high levels of questionable performance validity.

This is the first identified study of internal consistency of ImPACT scores. Assessing within-test reliability (internal consistency) on ImPACT was a challenge, as the ImPACT test generates a variety of score types and ranges to measure a variety of constructs. When looking at internal consistency in ImPACT subtest scores, the reliability between test administration procedures was not significantly different. The low reliability of the speed factor is noteworthy but may be due in part to the fewer items used in analysis.

There are a number of conclusions that can be drawn from these data. A primary conclusion is the high number of cases with questionable performance validity suggests that obtaining acceptable baseline data may be a challenge and makes the time and personnel required for baseline testing difficult to justify. Echemendia and colleague's review (2012) provides preliminary suggestion that baseline testing may not be any more efficient at identifying postconcussion deficits in cognition or balance than using only post-injury data and normative standards, although Schatz and Robertshaw (2014) found baseline comparison to be a more effective way to identify cognitive deficits in above average athletes.

Another important conclusion is that baseline (and post-injury) test interpretation is sufficiently complex and nuanced to require interpretation by a neuropsychologist. In the context of potentially suboptimal effort on baseline, using baseline data in a vacuum, without context or analysis, creates an increased risk of returning an athlete to play before they are fully recovered.

Finally, the combination of strong reliability in the face of high levels of performance invalidity raises some interesting possibilities. While the concept of embedded performance validity indicators is well-established, there is no proof of invalid effort in this sample, just indication of it. However, an alternative explanation would be that this sample of athletes were reliably performing at a level that was lower than their best effort would produce, and in effect, they were "sandbagging" a little consistently throughout the test. Based on anecdotal discussions with athletes, this is indeed a possibility.

Limitation

There were several limitations to the study, with the primary limitation being that the study design did not include randomized group assignment. ImPACT no longer reports race or ethnicity, so analysis was not conducted using that demographic variable. While there was good information about the administration protocol and environment for the standardized administration group, little information was available for the comparison group, thus limiting the inferences that can drawn. Finally, the use of internal consistency reliability (Cronbach's alpha) was somewhat atypical in that this metric is typically used to assess the test and not the respondents. However, given our hypothesis, alpha was used in this manner on the assumption that if some respondents were performing in a variable manner, this consistency metric (Cronbach's alpha) would reflect that pattern. Of course, the standard view also holds: that the internal consistency of the test factor is good.

Future Directions

An additional way to assess the validity of baseline data is to compare it to post-injury data. If an athlete is able to perform statistically better on post-injury testing (beyond practice effects and regression to the mean), it might suggest that their effort on the baseline test was suspect. This could provide a way to identify baselines that were completed without best effort and study suboptimal effort without using a simulator design, which has inherent limitations.

Finally, while the standardized protocol for Group 2 provided athletes the opportunity to identify if they were not feeling well or were short on sleep (and were rescheduled for later testing if identified), there were several athletes who endorsed less than 6 hr of sleep and had high symptoms scores on ImPACT, despite denying this to the examiners before testing. The findings on the effects of sleep are mixed (Mihalik et al., 2013; Silverberg, Berkner, Atkins, Zafonte, & Iverson, 2016), but some research (McClure, Zuckerman, Kutscher, Gregory, & Solomon, 2014) demonstrated that high school and college-aged athletes who slept less than 7 hr the night before baseline testing had lower scores on Verbal Memory, Visual Memory and Reaction Time composites. College athletes may be particularly vulnerable as they transition into college and complete the rigorous (often week-long) entrance process to college athletics. Analysis of baseline scores when athletes are only tested when they are feeling well and have been getting sufficient sleep may be telling in the elusive search for "best effort" on baseline testing.

Conflict of Interest — None

References

Barth, J., Alves, W., Ryan, T., Macciocchi, S., Rimel, R., Jane, J., & Nelson, W. (1989). Mild head injury in sports: neuropsychological sequelae and recovery of function. In H. Levin, H. Eisenberg, & A. Benton (Eds.), *Mild head injury* (pp. 257–275). New York, NY: Oxford University Press.

Binder, L., Iverson, G., & Brooks, B. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24*, 31–46. doi: 10.1093/arclin/acn001.

Broglio, S., Cantu, R., Gioia, G., Guskiewicz, K., Kutcher, J., Palm, M., et al.. (2014). National Athletic Trainers' Association position statement: Management of sport.

Echemendia, R., Bruce, J., Bailey, C., Sanders, J., Arnett, P., & Vargas, G. (2012). The utility of post-concussion neuropsychological data in identifying cognitive change following sports-related MTBI in the absence of baseline data. *The Clinical Neuropsychologist*, *26*, 1077–1091. doi: 10.1080/13854046. 2012.721006.

Erdal, K. (2012). Neuropsychological testing for sports-related concussion: How athletes can sandbag their baseline testing without detection. *Archives of Clinical Neuropsychology*, *27*, 473–479. doi: 10.1093/arclin/acs050.

Gaudet, C., & Weyandt, L. (2016). Immediate Post-Concussion and Cognitive Testing (ImPACT): A systematic review of the prevalence and assessment of invalid performance. *The Clinical Neuropsychologist*, *4046* (*1*), 1–16. doi: 10.1080/13854046.2016.1220622.

Harmon, K., Drezner, J., Gammons, M., Guskiewicz, K., Halstead, M., Herring, S., et al. (2013). American Medical Society for Sports Medicine position statement: Concussion in sport. *British Journal of Sports Medicine*, *47*, 15–26. doi: 10.1136/bjsports-2012-091941.

Higgins, K. (2015). *Sandbagging on the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) in a high school athlete population.* Doctoral dissertation, Forest Institute.

Higgins, K., Denney, R., & Maerlender, A. (2017). Sandbagging on the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) in a high school athlete population. *Archives of Clinical Neuropsychology*, *32*, 259–266. doi: 10.1093/arclin/acw108.

ImPACT Applications, Inc. (2011). Technical manual: Online ImPACT 2007–2012. Retrieved through email to info@impacttest.com.

Iverson, G., & Schatz, P. (2015). Advanced topics in neuropsychological assessment following sport-related concussion. *Brain Injury*, *29*, 263–275. doi: 10. 3109/02699052.2014.965214.

Larrabee, G. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*, 625–631. doi: 10.1017/S1355617712000240.

Larrabee, G. (2016). *Historical, conceptual and empirical factors in performance and symptom validity assessment*. Presented at the National Academy of Neuropsychology 36th Annual Conference, Seattle, WA.

Mayers, L., & Redick, T. (2012). Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues. *Journal of Clinical and Experimental Neuropsychology*, *34*, 235–242. doi: 10.1080/13803395.2012.667790.

McClure, D., Zuckerman, S., Kutscher, S., Gregory, A., & Solomon, G. (2014). Baseline neurocognitive testing in sports-related concussions: The importance of a prior night's sleep. *The American Journal of Sports Medicine*, *42*, 472–478. doi: 10.1177/0363546513510389.

McCrory, P., Makdissi, M., Davis, G., & Collie, A. (2005). Value of neuropsychological testing after head injuries in football. *British Journal of Sports Medicine*, *39*, i58–i63. doi: 10.1136/bjsm.2005.020776.

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

Mihalik, J., Lengas, E., Register-Mihalik, J., Oyama, S., Begalle, R., & Guskiewicz, K. (2013). The effects of sleep quality and sleep quantity on concussion baseline assessment. *Clinical Journal of Sports Medicine*, *23*, 343–348. doi: 10.1097/JSM.0b013e318295a834.

Moser, R., Schatz, P., Neidzwski, K., & Ott, S. (2011). Group versus individual administration affects baseline neurocognitive test performance. *The American Journal of Sports Medicine*, *39*, 2325–2330. doi: 10.1177/0363546511417114.

Moser, R. S., Schatz, P., & Lichtenstein, J. D. (2013). The importance of proper administration and interpretation of neuropsychological baseline and postconcussion computerized testing. *Applied Neuropsychology: Child*, *4*, 41–48. doi: 10.1080/21622965.2013.791825.

Pallant, J. (2005). *SPSS survival manual*. Crows Next NSW: Allen & Union. Piland, S., Ferrara, M., Macciocchi, S., Broglio, S., & Gould, T. (2010). Investigation of baseline self-report concussion symptom scores. *Journal of Athletic Training*, *45*, 273–279. doi: 10.4085/1062-6050-45.3.273.

Randolph, C. (2011). Baseline neuropsychological testing in managing sport-related concussion: Does it modify risk? *Current Sports Medicine Reports*, *10*, 21–26. doi: 10.1249/JSR.0b013e318207831d.

Schatz, P. (2010). Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *The American Journal of Sports Medicine*, *38*, 47–53. doi: 10.1177/0363546509343805.

Schatz, P., & Glatts, C. (2013). "Sandbagging" baseline test performance on ImPACT, without detection, is more difficult than it appears. *Archives of Clinical Neuropsychology*, *28*, 236–244. doi: 10.1093/arclin/act009.

Schatz, P., & Robertshaw, S. (2014). Comparing post-concussive neurocognitive test data to normative data presents risks for under-classifying "above average" athletes. *Archives of Clinical Neuropsychology*, *29*, 625–632. doi: 10.1093/arclin/acu041.

Silverberg, N., Berkner, P., Atkins, J., Zafonte, R., & Iverson, G. (2016). Relationship between short sleep duration and preseason concussion testing. *Clinical Journal of Sport Medicine*, *26*, 226–231. doi: 10.1097/JSM.0000000000000241.

Vaughan, C., Gerst, E., Sady, M., Newman, J., & Gioia, G. (2014). The relation between testing environment and baseline performance in child and adolescent concussion assessment. *The American Journal of Sports Medicine*, *42*, 1716–1723. doi: 10.1177/0363546514531732.