Sandbagging on the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) in a high school athlete population

Kathryn L. Higgins
University of Nebraska-Lincoln, khiggins@unl.edu

Robert L. Denny
Neuropsychological Associates of Southwest Missouri

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

Published in *Archives of Clinical Neuropsychology* 32 (2017), pp 259-266.
doi 10.1093/arclin/acw108

Accepted 28 November 2016; published online 19 December 2016. PMID: 28431031

Sandbagging on the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) in a high school athlete population

Kathryn L. Higgins,[1] Robert L. Denney,[2] and Arthur Maerlender[1]

1 Center for Brain, Biology and Behavior, University of Nebraska-Lincoln, Lincoln, NE, USA
2 Neuropsychological Associates of Southwest Missouri, Springfield, MO, USA

Corresponding author — K. L. Higgins, Center for Brain, Biology and Behavior, University of Nebraska-Lincoln, C76, East Stadium, PO Box 880156, Lincoln, NE 68588, USA; email khiggins@unl.edu

Abstract

The Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) is a computerized neuropsychological test battery commonly used to assess cognitive functioning after a concussion. It is recommended that application of ImPACT utilizes a baseline administration so athletes have an individualized baseline with which to compare post-injury results should they sustain a concussion. It has been suggested that athletes may provide sub-optimal effort, called "sandbagging," in order to return to their baseline cognitive scores, and thus to play, more quickly. This research examines ImPACT baseline scores when high school athletes were asked to attempt to "sandbag," and compares those scores with scores obtained when they were asked to give their "best effort." Fifty-four high school student athlete volunteers participated in the study. In contrast to previous research that just looked at the cut-score invalidity indicators built into ImPACT, this research developed a regression equation to predict sandbagging. A logistic regression equation developed with four variables that demonstrated the largest effect size between "best effort" and "sandbagged" baselines showed a 99.7% classification accuracy for the "best effort" and "sandbag" groups.

Keywords: neuropsychological testing, high school athletes, baseline, concussion, concussion assessment

Introduction

Approximately 300,000 sports-related concussion are estimated to occur in the United States among high school athletes every year (Faure & Pemberton, 2010), between 1.6 and 3.8 million in the general U.S. population annually (Broglio, Pontifex, O'Connor, & Hillman, 2009), and sports-related concussions account for about 20% of all head injuries received annually in the United States (Rosenbaum & Arnett, 2010). Research suggests that 20% of all high school football players receive at least one concussion in their high school career (Theriault, De Beaumon, Gosselin, Filippinni, & Lassonde, 2009). It is also assumed that prevalence statistics are underestimated due to lack of continuity in diagnostic criteria, lack of education, and athletes failing to report their symptoms (Faure & Pemberton, 2010).

Computerized neurocognitive testing (CNT) is commonly used to assess cognitive functioning. CNT, such as the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT), uses the individual athlete's baseline scores to assess post-injury neurocognitive function and improvement after injury across domains of processing speed, reaction time, and concentration and memory (Faure & Pemberton, 2010).

Baseline testing was first introduced in 1982 as a part of the University of Virginia prospective study of mild head injury in football (Barth et al., 1989). Baseline testing offers a reference point to guide return-to-play decisions. Baseline testing also allows individual athletes to serve as their own controls, rather than relying on normative standards to ascertain when a concussed athlete has returned to "normal" (Covassin, Stearne, & Elbin, 2008; Yard & Comstock, 2009). 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. Despite the importance of importance of reliable and valid baseline data, baselines are frequently not assessed (or not assessed sufficiently) for invalid results. In a 2009 survey of athletic trainers, 95% of those who responded endorsed using ImPACT for baseline testing, but only 55% examined baselines for valid results (Covassin, Elbin, Stiller-Ostrowski, & Kontos, 2009). The Technical Manual for ImPACT lists common causes of invalid test scores, including failure to read or understand directions, attention-deficit/hyperactivity disorder (ADHD), fatigue, athletes distracting each other, or left/right confusion (ImPACT Applications, Inc., 2011). Current recommendations for baseline testing on ImPACT include re-administering a baseline that is flagged as "invalid" by ImPACT after discussing the results with the athlete and ascertaining the cause of the invalid results (ImPACT Applications, 2011). Schatz and colleagues (2014) found that 6.3% of high school athletes received at least one invalid score on ImPACT baselines, but that 90% of athletes who received invalid results will obtain valid baseline results on a subsequent, re-administered baseline. Consensus statements from two of the major bodies in the field, the American Academy of Clinical Neuropsychology (AACN) and the National Academy of Neuropsychology (NAN), have identified standards for recommendations regarding the measurement of effort in neuropsychological testing and the identification of suboptimal effort and/or negative response bias (Heilbronner et al., 2009). AACN's consensus statement (2009) identified effort as "occurring on a continuum" (p. 1097). Therefore, identifying the range (ex. "sufficient" vs. "insufficient") within which an examinee's effort fell is as precise as the identification can be, particularly given the limited data points for effort in baseline testing.

Assessing effort within the context of baseline testing can be problematic. There may be limitations on time and facilities, as well as the availability of personnel who are appropriately trained to administer and interpret additional effort testing. Therefore, identifying indicators that are already a part of the testing used (i.e., ImPACT) is the most feasible way to assess an athlete's effort.

Poor effort may be less prevalent in post-injury testing in an athlete population due to the test results' pivotal role in athletes' return to play, but suboptimal effort can playa significant role in an athlete's performance at baseline testing. Bailey, Echemendia, and Arnett (2006) studied a collegiate athlete population and found that those who were classified as demonstrating suboptimal effort showed significantly greater improvement on several neurocognitive measures than those who demonstrated high effort week 1 week post-injury. The authors suggested that effort played a role in the athletes' cognitive performance, as they were able to score significantly better when they knew their return to play was riding on their test results. Poor effort on baseline testing creates an invalid benchmark for comparison at post-injury and may indicate recovered neuropsychological functioning prematurely. Embedded indicators of changes in scores are rendered useless, while the athlete is put at increased risk of re-injury when they are returned to play before full recovery.

Although the ImPACT test has several indicators of invalid test performance built into the test, Erdal (2012) found that 11 % of an ImPACT savvy collegiate athlete population was able to successfully sandbag a baseline ImPACT test without activating the invalidity indicators. Successful faking was unaffected by gender, sport, or number of previous concussions. There were significant differences between the athletes' legitimate baselines and "sandbagged" baselines on all variables, except Reaction Time and Three Letters Correct. In contrast, Schatz and Glatts (2013) determined that in an ImPACT-naive, non-athlete sample of college students, 30% of uncoached participants and 35% of coached participants were able to sandbag without being caught by the ImPACT invalidity indicators.

Schatz and Glatts (2013) found that their group providing "best effort" and those who were coached on how to sandbag scored similarly on Reaction Time composite and Impulse Control composite, and coached and uncoached sandbaggers scored similarly on Verbal Memory Composite, Visual Memory Composite, and Total Symptoms Score. Schatz and Glatts (2013) found that a score of less than 22 on Word Memory Correct Distractors (Immediate + Delay) accurately identified 95% of uncoached sandbaggers and 100% of coached sandbaggers. They also found a cut-score of 16 on Design Memory Correct Distractors (Immediate + Delay) was useful for identifying sandbagging. It accurately identified 90% of uncoached sandbaggers and 95% of coached sandbaggers, although it also incorrectly identified 20% of the "best effort" group as sandbagging, which decreases the utility of the measure.

The current research study attempts to expand understanding of suboptimal effort on ImPACT baseline testing. It is an expansion of previous research, as prior research on sandbagging has used college athletes/students, while this study utilized a sample of high school athlete participants who had prior exposure to ImPACT (ImPACT savvy). In addition, this study attempts to identify a combination of variables for more accurately identifying those high school athletes who are sandbagging (providing insufficient effort) on the ImPACT. Of note, the label "best effort" is used throughout this study to specify that particular administration parameter. This does not imply that all athletes provided their most superior effort, as that would be nearly impossible to quantify and measure in this context. This is simply the label that was used to differentiate between the administration in which athletes were asked to "give their best effort" (and provided adequate effort as best as that is able to be determined) and the sandbagged administration. This terminology is also in line with other published studies in this area of research. Quotations will be used to indicate that this is an administration label, rather than a statement of the type of effort provided.

Hypotheses

(1) Baseline scores of high school athletes told to intentionally sandbag baseline tests will be significantly worse than baseline scores of those same athletes when told to provide their "best effort."

(2) A small number of subtest scores will demonstrate high classification accuracy for the "best effort" and "sandbagging" groups.

(3) A composite performance validity score (utilizing those subtests with the greatest classification accuracy) will show the highest classification accuracy when compared with those individual subtests.

Materials and Methods

Participants

Participants were 66 high school athletes from high schools in the rural Midwest. All athletes had previously completed an ImPACT baseline through their school's concussion management program. Participants were recruited through their various athletic teams, using a short recruitment speech given at the end of team meetings and signs in the locker room areas. The athletic trainer also played an active role in athlete recruitment, and placed recruitment signs in the training room and locker rooms. As an incentive for participation, athletes were provided with food after testing, and were entered into a drawing for a gift card. All of an individual's data were excluded if the computer did not save both of their baselines or they did not complete two baselines

(n = 4); or if they indicated on the Additional Demographics that they had not previously completed an ImPACT baseline (n = 5); or if their "best effort" baseline was flagged as invalid by the ImPACT Invalidity Indicators (n = 3). The final total sample consisted of 54 athletes.

All participants and their legal guardians signed an informed consent that was approved by the Institutional Review Board. Formal Institutional Review Board Approval was obtained for this project by the author's institution and data collection continued until sufficient data had been collected.

Materials

All participants were administered ImPACT Version 2.1. 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.

ImPACT Applications, Inc. Technical Manual (2011) reports that interclass correlation coefficients (ICCs) for composite scores in a group of high school athletes tested I year apart ranged from .57 to .86. Schatz (2010) found that baseline ImPACT tests given to collegiate athletes approximately 2 years apart had ICC's ranging from .46 to .74 for the composites and .44 for the symptoms scale.

Procedure

Baseline testing was completed in the computer laboratory on the participants' campus. Per ImPACT Application, Inc. administration instructions, each testing cohort ranged from 2 to 14 athletes. All athletes completed two administrations of ImPACT; once with instructions to give their "best effort" (BE) and once with directions to provide suboptimal effort called "sandbagging" (SB). Administration order was counter balanced in an attempt to control for order effects. Each testing cohort was randomly assigned to either BE/SB administration (completed BE baseline, then SB baseline) or SB/BE administration (completed SB baseline, then BE baseline). Baselines were administered back-to-back. Instructions scripts for both administrations are in Appendix.

Analyses

Descriptive statistics for participants were analyzed, including statistics for previous concussions. Data were assessed for normality and outliers, as well as for order effects. Independent samples *t-tests* were run to examine between-group differences on various demographic and testing variables for both BE and SB baselines. Internal consistency reliability of the 26 subtest scores was assessed for the BE baselines to assess the stability of the scores. Test-retest reliability would have been the preferable metric, but given the differing administration conditions under which the baselines were taken, that was not an appropriate metric.

Independent samples *t*-tests were also run to examine group differences between administration order, although the crossover nature of the experimental design makes group differences irrelevant to the overall statistical findings. Paired samples *t*-tests (or Wilcoxon Signed Rank Tests for non-normally distributed composite scores) were then run to examine differences between BE and SB baselines. Bonferroni correction was used to control for Type 1 error.

Argesti's (2007) recommendation that independent variables entered in to a logistic regression analysis be limited to 1 for every 10 subjects, a limit of 5 independent variables was set, based on our samples size of 54. Because of the generally large effect sizes and large number of variables, those subtest scores with the largest effect sizes ($d > 2.00$) were initially identified for the analysis. To decrease multicollinearity, scores were excluded if they were created using other subtest scores (i.e., Word Memory Total Percent Correct = Word Memory Learning Percent Correct + Word Memory Delayed Percent Correct), which only left three scores. The item with the next largest effect size under 2.00 — Design Memory Total Percent Correct (DM) — was pulled for the logistic regression equation in order to utilize the greatest number of variables, given the number of participants. A forced-entry logistic regression analysis was conducted with Test Type (BE vs. SB) as the dependent variable. Independent variables included in the model were Word Memory Learning Percent Correct (WM LP), Word Memory Delay Memory Percent Correct (WM DM), Design Memory Total Percent Correct (DM), and X's and O's Total Correct (Interference) (XO). The unstandardized beta (β) value for each variable of the logistic regression analysis was combined into a predictive regression equation using Euler's constant (e = 2.718). *"e"* represents a mathematical constant that is the base of the natural logarithm. The constant *"e"* is raised to an exponent, which is created by adding the coefficient for the constant to the β-value for each variable times the participant's score on the variable. This can be used to create a single variable to predict sandbagging using a single athlete's test scores. All variables entered in the logistic regression were used in the equation to reduce bias and decrease the risk of Type I error.

The individual subtest scores used in the logistic regression, as well as the single variable output of the regression equation, were then assessed for sensitivity and specificity using receiver-operator characteristics.

Results

Participants' average age was 16.81 (SD = 0.87) years; a majority were either sophomores (n = 15; 27.8%) or juniors (n = 27; 50.0%). A majority of the participants identified as White (n = 52; 96.3%) and men (n = 33; 61.1 %). A majority of the participants (n = 27; 50.0%) were participating in a high school football program. The remainder were participating in the high school programs of basketball (n = 15; 27.8%), volleyball (n = 8; 14.8%), or softball (n = 4; 7.4%).

Due to differing group sizes and random selection of cohort assignment to administration, 34 athletes (63%) completed the BE administration first, then the SB administration, and 20 athletes (37%) completed SB, then BE administration. Examination of order effects found on BE administration Visual Motor Speed composite, those athletes who completed BE administration second performed faster (M = 45.90, SD = 5.19) than those who completed BE administration first (M = 41.81, SD = 5.22; $t(52) = -2.79$, p = .007). Significant differences were also found on SB administration Verbal Memory composite, with those who completed SB administration first (M = 65.25, SD = 9.13) remembering a greater number of words than those who completed the SB administration second (M = 57.68, SD = 14.63; t (52) = −2.08, p = .042).

The 26 subtest scores from the BE baselines had a Cronbach's α = 0.76, demonstrating an acceptable level of internal consistency.

Significant differences were between all composite scores in the direction expected (i.e., SB reaction time scores were higher than BE, SB memory scores were lower than BE). **Table 1** lists the means, standard deviations, *t*-scores, and effect sizes for each composite score.

There were also significant differences between all BE and SB subtest scores with the exception of Symbol Match Average Correct Reaction Time (Visible) and Symbol Match Average Correct Reaction Time (Hidden). All significantly different subtest scores demonstrated medium to large effect sizes, as well (d = −0.63 to 3.52).

The four variables selected for the logistic regression equation, their means, standard deviations, *t*-scores, and effect sizes are listed in **Table 2**.

Table 1. Differences in composite scores between best effort "BE" (n = 34) and Sandbagged "s" (n = 20) baselines

Composite	*BE, M(SD)*	*SB, M(SD)*	*t-score*	*Effect size (d)*
Verbal Memory[a]	87.70 (8.01)	60.48 (13.30)	14.57	2.48
Visual Memory[a]	77.63 (13.69)	57.61 (11.89)	9.05	1.56
Visual Motor Speed[a]	43.33 (5.53)	32.12 (8.53)	9.02	1.56
Reaction Time[b]	Md = 0.054	Md = 0.65	z = −5.16	r = 0.50
Impulse Control[a]	7.32 (4.59)	27.28 (19.02)	−7.96	−1.44
Total Symptom Score[b]	Md = 2.00	Md = 32.00	z = −5.78	r = 0.56

a. Paired samples *t*-test. $t(53)$, $p < .001$.
b. Wilcoxon Signed Rank Test, $p < .001$.

Table 2. Differences in subtest scores used in logistic regression

Subtest	*BE, M(SD)*	*SB, M(SD)*	*t-score*	*Effect size (d)*
WMLP	97.22 (4.08)	65.46 (12.09)	17.90	3.52
WMDP	92.80 (6.74)	63.53 (12.38)	13.78	2.75
DM	84.12 (13.06)	61.50 (10.37)	11.93	1.92
XO	116.96 (7.34)	87.43 (17.56)	12.53	2.19

WM LP = Word Memory Learning Percent Correct
WM DP = Word Memory Delayed Memory Percent Correct
DM = Design Memory Total Percent Correc
XO = X's and O's Total Correct (interference)

Logistic Regression Analysis

Analysis of multicollinearity demonstrated that levels were within acceptable limits for the included variables (Tolerance >.10 and variance inflation factor <10; Pallant, 2005). The full logistic regression model was statistically significant, χ^2 (4, N = 108) = 133.32, p < .001, indicating that the model was able to accurately differentiate SB from BE scores. The model explained between 70.9% (Cox & Snell R Square) and 94.5% (Nagelkerke R Square) of the variance in effort, and correctly classified 97.2% of the athletes' effort on baseline testing. None of the individual subtests offered a statistically significant contribution to the model, with p-levels from .07 to .18 and odds ratio (Exp [B]) from 0.83 to 0.88, as displayed in **Table 3**.

Based on the results of the logistic regression, the following equation was determined to predict sandbagging:

$$\frac{e^{(56.74 - (0.15 \times \text{WM LP}) - (0.18 \times \text{WM DM}) - (0.13 \times \text{DM}) - (0.17 \times \text{XO}))}}{1 + e^{(56.74 - (0.15 \times \text{WM LP}) - (0.18 \times \text{WM DM}) - (0.13 \times \text{DM}) - (0.17 \times \text{XO}))}}$$

Receiver Operating Characteristic Analysis

Receiver operating characteristic curves were assessed for each independent variable included in the logistic regression, as well as the output variable of the logistic regression equation. Sensitivity, specificity, and area under the curve (AUC) for various cut-scores are

Table 3. Variables in logistic regression predicting sandbagging on baseline ImPACT

Variable	*B*	*SE*	*Wald*	*df*	*Sig*	*Exp (B)*
WMLP	−0.15	0.08	3.29	1	0.07	0.86
WMDM	−0.18	0.11	2.67	1	0.10	0.83
DM	−0.13	0.09	2.30	1	0.13	0.88
XO	−0.17	0.13	1.83	1	0.18	0.84
Constant	56.74	24.62	5.19	1	0.02	4.38

WM LP = Word Memory Learning Percent Correct
WM DP = Word Memory Delayed Memory Percent Correct
DM = Design Memory Total Percent Correct
XO = X's and O's Total Correct (interference)

Table 4. Classification accuracy. sensitivity and specificity of variables and equation

Variable	AUC	Cut-score	Sensitivity (%)	Specificity (%)
WM LP	.981	≤90%	96.3	94.4
WM DM	.976	≤81%	94.4	94.4
DM	.896	≤64.75%	61.1	89.9
XO	.961	≤107.5	90.7	89.9
Logistic regression equation	.997	≤0.23	100.0	90.7

WM LP = Word Memory Learning Percent Correct
WM DP = Word Memory Delayed Memory Percent Correct
DM = Design Memory Total Percent Correct
XO = X's and O's Total Correct (interference)

recorded in **Table 4**. Per Boone (2007), cut-scores with 90% specificity are recommended to decrease the rate of false positives for sandbagging. All AUC values fell in the "outstanding" range (Hosmer, Lemeshow, & Sturdivant, 2013), with the exception of Design Total Memory Percent Correct, which fell in the "excellent" range (Hosmer, Lemeshow, & Sturdivant, 2013) and shown in **Table 4**.

The full equation demonstrates the best overall classification accuracy with AUC = 0.997 ("outstanding" range). A cut-score of ≤.23 from the logistic regression equation demonstrated 90.7% specificity and 100.0% sensitivity.

Discussion

This study was successful in identifying variables that differentiated poor test effort from sufficient effort. This study confirmed previous research that has shown that nearly exclusively, ImPACT savvy, high school-aged athletes' performances were poorer on baseline tests when they were asked to "sandbag" than when they were asked to give their "best effort." These differences were seen on both composite and subtest scores. In fact, athletes attempting to "sandbag" had a poorer performance at a statistically significant level on every score-both composite and subtest — with the exception of two (Symbol Match Average Correct Reaction Time for the Hidden and Visible trials).

Further, the study successfully demonstrated that a subset of subtest scores could differentiate effort levels over and above individual subtests. Candidate subtests were combined in an equation to directly compare the combined subtest scores with the individual subtests for sensitivity and specificity. Indeed, the equation using WM LP, WM DM, DM, and XO accurately classify 99.7% of the athletes' level of effort. A cut-score of ≤.23 from the logistic regression equation demonstrated 90.7% specificity and 100.0% sensitivity.

In order to demonstrate the utility of the equation, a single athlete's data from the sample are shown below and its application using the equation. This athlete's scores on their "sandbag" administration were as follows: WM LP = 75% , WM DM = 58%, DM = 54.5%, XO = 102. This baseline was not flagged as invalid (i.e., Baseline ++) by the ImPACT test. The application of this athlete's scores to the equation is as follows:

$$\frac{e^{(56.74 - (0.15 \times 75) - (0.18 \times 58) - (0.13 \times 54.5) - (0.17 \times 102))}}{1 + e^{(56.74 - (0.15 \times 75) - (0.18 \times 58) - (0.13 \times 54.5) - (0.17 \times 102))}}$$

The output of the equation when using this athlete's data is 1.00, which falls above the 0.23 cutoff established.

In contrast to the equation, this study found that the ImPACT validity indicators correctly identified approximately 65% of the baseline tests that athletes had been asked to "sandbag." This finding is similar to the results by Schatz and Glatts (2013), who found that the ImPACT invalidity indicators identified 65-70% of their ImPACT-naïve, non-athlete college student sample.

The multiple large effect sizes seen between Word Memory scores on BE and SB baselines are likely due to the relatively easy nature of the Word Memory task. In the current sample, high school athletes were able to identify 97% of words they had seen or had not seen immediately and 92% after a delay. Because athletes providing best effort accurately identify so many of the words, it becomes obvious when they sandbag, even though they miss only a few words. The high percentage of words that athletes identify when providing BE makes this task difficult to sandbag subtly enough to not create an obvious difference between those athletes who are providing best effort and those who are not. In contrast, when athletes provided BE on Design Memory, they were only able to identify 84% of the designs that they had seen or not seen immediately with 82% accuracy after a delay. Like Schatz and Glatts (2013), Design Memory was found to have slightly less utility than Word Memory.

Erdal (2012) found that the athletes in her sample who were successfully able to sandbag without being caught had higher Word Memory Learning Percent Correct scores on their BE baselines. The current sample also demonstrated larger differences in mean scores between BE and SB baselines across the Word Memory Module, which would suggest that athletes who would normally correctly identify almost all of the words, still fall above the range of detection when they sandbag. In Erdal's (2012) study, a Verbal Memory composite of ≤70% identified 73% of the sandbaggers. In the current sample, Word Memory Learning Percent Correct and Word Memory Delay Memory Correct demonstrated the greatest ability to classify the BE and SB baselines. These two scores combine to form part of the Verbal Memory Composite, suggesting that seeming inability to learn and remember words is highly indicative of sandbagging.

Limitations

There are several limitations to this study that should be noted. One of the most significant limitations was the homogeneity of the group. An overwhelming majority of the group was white, Midwestern, from a small town, with no reported diagnoses. All diagnoses (or lack thereof) were based on self-report, so the reliability of that information may be questionable. These metrics may not apply to those with known learning or attention deficits, or other notable health histories should not be assessed with these metrics. The test environment was relatively controlled, but there were between 4 and 14 athletes per group, which may have introduced variability into the scores (Moser et al., 2011). The baseline tests were administered back-to-back, which is not traditional administration, but is not contraindicated by any ImPACT instructional material. Finally, the regression-based predictive formula requires independent cross-validation before it can be shown to be shown as clinical useful for determining sandbagging in baseline testing among high school athletes. Only athletes with prior exposure to ImPACT were utilized in this sample, so generalizability to "ImPACT-naïve" athletes may be limited, as ImPACT composite scores have been demonstrated to change with multiple exposures to the test (Maerlender et al., 2016).

In addition to cross-validation of the predictive formula among high school athletes, the utility of the formula should be researched among collegiate and professional athletes. Because one of the most effective predictors of sandbagging is Word Memory, it is vital that research continues to be done on the performance of athletes with diagnoses that could potentially affect word learning and memory (such as dyslexia and learning disabilities). It is not uncommon for high school and collegiate athletes to have these conditions; consequently, future research should also include a focus on these clinical groups.

Funding — This work was supported by a Research Grant from Forest Institute of Professional Psychology.

Conflict of Interest — None declared.

Acknowledgments — A portion of this data was presented at the Sports Neuropsychology Concussion Symposium 2016.

Appendix

Group Scripts

BE/SB Group Script

Thank you for participating in my research. Today, I will be asking you to take a test, the Immediate Post-concussion Assessment and Cognitive Testing, or the ImPACT. The baseline (or initial) ImPACT is given to an athlete before they begin a season, so that the trainer or doctor will know the athlete's baseline ability should the athlete receive a concussion. This is not an intelligence test and there are some tests that will be very easy and some that will be more difficult. The test has ways of telling is someone is not giving their best effort. I will be asking to ask you to take the test twice.

The first time that you take the test, please give your best effort. Please concentrate and complete the test as quickly and accurately as possible. The second time I would like you to give less than your best effort. Remember that if you perform too poorly, the test will pick up on that and indicate that the baseline is invalid. I would like you to try and "trick the system" by performing in the area between your best effort and so poorly that the test is invalid due to poor effort. Remember to also consider how you might respond to the symptoms ratings differently this time around.

While you may have to take this test to assess a real concussion at some point in the future, these two administrations will not be used for anything other than my research. If you have any questions during the two administrations of the test, please raise your hand. After you have finished each administration, please sit quietly as to not disturb those around you.

Thank you

SB/BE Group Script

Thank you for participating in my research. Today, I will be asking you to take a test, the Immediate Post-concussion Assessment and Cognitive Testing, or the ImPACT. The baseline (or initial) ImPACT is given to an athlete before they begin a season, so that the trainer or doctor will know the athlete's baseline ability should the athlete receive a concussion. This is not an intelligence test and there are some tests that will be very easy and some that will be more difficult. The test has ways of telling is someone is not giving their best effort. I will be asking to ask you to take the test twice.

The first time that you take the test, I would like you to give less than your best effort. Remember that if you perform too poorly, the test will pick up on that and indicate that the baseline is invalid. I would like you to try and "trick the system" by performing in the area between your best effort and so poorly that the test is invalid due to poor effort. Remember to also consider how you might respond to the symptoms ratings differently this time around.

The second time, please give your best effort. Please concentrate and complete the test as quickly and accurately as possible.

While you may have to take this test to assess a real concussion at some point in the future, these two administrations will not be used for anything other than my research. If you have any questions during the two administrations of the test, please raise your hand. After you have finished each administration, please sit quietly as to not disturb those around you.

Thank you

References

Agresti. A. (2007). *An introduction to categorical data analysis.* Hoboken. NJ: John Wiley & Sons.

Bailey, C., Echemendia, R., & Arnett, P. (2006). The impact of motivation on neuropsychological performance in sports-related mild traumatic brain injury. *Journal of the International Neuropsychological Society.* 12, 475-484.

Barth, J., Alves, W., Ryan, T., Macciocchi, S., Rimel, R., Jane, J., et al. (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: Oxford University Press.

Boone, K. (2007). A reconsideration of the Slick et al. (1999) criteria for malingered neurocognitive dysfunction. In Boone. K. B. (Ed.) *Assessment of Feigned Cognitive Impairment* (pp. 29–49). New York. NY: The Guilford Press.

Broglio, S., Pontifex, M., O'Connor, P., & Hillman, C. (2009). The persistent effects of concussion on neuroelectric indices of attention. *Journal of Neurotrauma,* 26, 1463-1470.

Covassin, T., Elbin, R., Stiller-Ostrowski, J., & Kontos, A. (2009). Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) practices of sports medicine professionals. *Journal of Athletic Training,* 44, 639–644.

Covassin, T., Stearne, D., & Elbin, R. (2008). Concussion history and postconcussion neurocognitive performance and symptoms in collegiate athletes. *Journal of Athletic Training,* 43, 119-124.

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

Faure, C., & Pemberton, C. (2010). Concussion and the young athlete: Critical management strategies. *The Journal of Physical Education Recreation and Dance,* 81, 19-26.

Heilbronner, R., Sweet, J., Morgan, J., Larrabee, G., & Millis, S., Conference Participants. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist,* 23, 1093-1129.

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression* (3rd ed). Hoboken. NJ: John Wiley & Sons, Inc.

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

Maerlender, A. C., Masterson, C. J., James, T. D., Beckwith, J., Brolinson, P. G., Crisco, J., et al. (2016). Test-retest, retest, and retest: Growth curve models of repeat testing with Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT). *Journal of Clinical and Experimental Neuropsychology,* 38, 869-874.

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 (11), 2325-2330.

Pallant, J. (2005). *SPSS survival manual* (2nd ed). NY: Open University Press.

Rosenbaum, A., & Arnett, P. (2010). The development of a survey to examine knowledge about and attitudes toward concussion in high-school students. *Journal of Clinical & Experimental Neuropsychology,* 32, 44–55.

Schatz, P. (2010). Long-term test-retest reliability of baseline cognitive assessment using ImPACT. *American Journal of Sports Medicine,* 38(1), 47-53.

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.

Schatz, P., Kelley, T., Ott, S., Solomon, G., Elbin, R., Higgins, K., et al. (2014). Utility of repeated assessment following invalid baseline performance. *Journal of Athletic Training,* 49, 659-664.

Theriault, M., De Beaumon, L., Gosselin, N., Filippinni, M., & Lassonde, M. (2009). Electrophysiologial abnormalities in well-functioning multiple concussed athletes. *Brain Injury,* 32, 899–906.

Yard, E., & Comstock, D. (2009). Compliance with return to play guidelines following concussion in US high school athletes, 2005-2008. *Brain Injury, 23,* 888-898.