OXFORD
UNIVERSITY PRESS

Archives
of
CLINICAL
NEUROPSYCHOLOGY

One-Month Test–Retest Reliability of the ImPACT Test Battery

Philip Schatz*, Charles S. Ferris

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

*Corresponding author at: Department of Psychology, Saint Joseph's University, 5600 City Avenue, Post Hall 222, Philadelphia, PA 19131, USA.
Tel.: +1-610-660-1804; fax: +1-610-660-1819.
E-mail address: pschatz@sju.edu (P. Schatz).

Accepted 18 April 2013

Abstract

In order to better understand the test–retest reliability of the ImPACT test battery, 25 undergraduate students completed two ImPACT tests across a time frame of 4 weeks between assessments. Participants had not previously completed ImPACT and had no history of concussion. Pearson's correlation coefficients (*r*) and intraclass correlation coefficients (ICCs) were as follows: Verbal Memory = .66/.79 (r/ICC), Visual Memory = .43/.60, Visual Motor Speed = .78/.88, Reaction Time = .63/.77, and Total Symptoms = .75/.81. Dependent sample *t*-tests revealed significant improvement on only Visual Motor Speed composite scores. Reliable Change Indices showed a significant number of participants fell outside 80% and 95% confidence intervals for only Visual Motor Speed scores (but no other indices), whereas all scores were within 80% and 95% confidence intervals using regression-based measures. Results suggest that repeated exposure to the ImPACT test may result in significant improvements in the physical mechanics of how college students interact with the test (e.g., performance on Visual Motor Speed), but repeated exposure across 1 month does not result in practice effects in memory performance or reaction time.

Keywords: ImPACT; Concussion; Neuropsychological testing; Mild head injury; Baseline screening

Introduction

The utility of the ImPACT computer-based neuropsychological test, for the purpose of diagnosis, assessment, and management of sports-related concussion, has been a topic of debate in the literature (Lovell, 2006; Mayers & Redick, 2012; Randolph, 2011; Randolph, Lovell, & Laker, 2011; Randolph, McCrea, & Barr, 2005, 2006; Schatz, Kontos, & Elbin, 2012). A central part this debate has been the reliability of this measure, which has been documented over a period of time ranging from days (Iverson, Lovell, & Collins, 2003) to months (Broglio, Ferrara, Macciocchi, Baumgartner, & Elliott, 2007) to years (Elbin, Schatz, & Covassin, 2011; Schatz, 2010). Given that reliability coefficients were lower across a 45- and a 50-day period (Broglio et al., 2007) than across 1-year (Elbin et al., 2011) and 2-year (Schatz, 2010) periods, the methodology used by Broglio et al., (2007) was questioned (Schatz, Kontos, et al., 2012). In Broglio's 2007 study, 34% participants were excluded, due to invalid data, which is considerably higher than documented incidences of: 2.5%–8.7% of high-school athletes (Schatz, Neidzwski, Moser, & Karpf, 2010; Schatz, Pardini, Lovell, Collins, & Podell, 2006), 5.1% of collegiate athletes (Schatz, 2010), and 5.0% of professional athletes (Solomon & Haase, 2008). Further, the remaining 66% of participants in the Broglio et al., (2007) study may have experienced fatigue, proactive, and retroactive interference (e.g., due to successive administration of multiple computerized test batteries), resulting in increased variability in the sample (Schatz, Kontos, & Elbin, 2012).

The current study was conducted in order to document the test–retest reliability across a time period consistent with an athlete completing a preseason baseline and sustaining a concussion within that same athletic season or academic semester. As a result, we examined the test–retest reliability of the ImPACT test battery, administered independent of any other tests, between assessments across a 4-week time period.

doi:10.1093/arclin/act034

Methods

Participants

Participants were 28 undergraduate students recruited from the psychology department's student research pool. Recruitment exclusion criteria included varsity athlete status, previous exposure to the ImPACT test, previous history of concussion, or invalid performance on the test. Twenty-eight participants completed an initial baseline test, but only 25 were included in the analyses as two participants failed to show for the second assessment, and one participant was excluded due to suspected invalid performance (e.g., scoring more than 3 *SD* below the normative mean on Visual Memory; *see* Lovell, 2007). Female participants (76%), outnumbered males (24%), and all participants reported English as their first language. No participants reported a history of: Attention Deficit Disorder/Hyperactive Disorder, Learning Disability, or treatment for headaches, migraines, seizures, or other neurologic or psychiatric illness.

Materials/Procedure

Saint Joseph's University's Institutional Review Board approved experimental procedures, and each participant provided informed consent. Participants completed the computerized ImPACT baseline test (online version) and returned 4 weeks later for the second assessment. Testing took place in a quiet laboratory setting and was conducted either individually or in pairs, supervised by the same individual (CF). Participants completed the "Baseline" version of ImPACT on the first testing session, and then the "Post-Injury 1" version on the second. These test versions are essentially identical, but incorporate different stimuli (e.g., for memory tasks) or placement of stimuli (e.g., for visual-motor tasks).

Analyses

Participants completing consecutive assessments comprised a within-subjects sample, allowing for the comparison between Times 1 and 2. Dependent measures included the Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time composite scores from ImPACT, and the Total Symptom scores.

Pearson's product–moment correlations (r) were calculated as a general measure of the strength of linear association between variables at Times 1 and 2. Pearson's r is considered a weak measure of test–retest reliability, in situations where coefficients are high and group means are similar, but there is considerable variation in individual scores from Times 1 to 2 (Rodgers & Nicewander, 1998). As such, intraclass correlation coefficients (ICCs), considered a better measure of association than Pearson's r (Wilk et al., 2002), were calculated as the primary indicator of test–retest reliability. ICC can distinguish those sets of scores that are merely ranked in the same order from test to retest from those that are not only ranked in the same order but are in low, moderate, or complete agreement (Chicchetti, 1994). The ICC model, "Two-Way Mixed" type "Consistency," was used, using "Average Measures" (Weir, 2005); ICC analyses also yield an Unbiased Estimate of Reliability, which reflects the consistency of the baseline assessments (Fleishman & Benson, 1987).

Reliable Change Indices (RCI; Jacobson & Truax, 1991) are 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 score would not be obtained as a result of the measurement error (Iverson, Sawyer, McCracken, & Kozora, 2001). We calculated and employed a modified RCI formula (Chelune, Naugle, Lüders, Sedlak, & Awad, 1993), which includes an adjustment for practice effects (*see* Barr, 2002, for a more detailed discussion). Finally, regression-based methods (RBMs) were applied to the data. In the RBM, the scores from the first assessment are placed into a regression analysis, using the score at Time 2 as the dependent variable, with the resulting equation providing an adjustment for the effect of the initial performance level, as well as controlling for any regression to the mean (McCrea et al., 2005). With this technique, regression equations can be built to predict a participant's level of performance on a neuropsychological instrument at retest from the initial testing (Crawford & Garthwaite, 2006). Following Bonferonni correction, the α level for dependent-samples analyses was set at $p < .01$.

Results

Pearson's correlations between baseline assessments ranged from 0.43 to 0.78. ICCs reflected higher reliability than Pearson's r, across all measures. Visual Motor Speed scores showed the most stability (mean ICC = 0.88; 0.72–0.95; lower and upper 95% confidence intervals), followed by Verbal Memory (0.79; 0.52–0.91), Total Symptom scores (0.81; 0.57–0.92), Reaction Time (0.77; 0.47–0.90), and Visual Memory (0.60; 0.08–0.84). Unbiased Estimates of Reliability were consistent with ICCs: Visual Motor Speed (0.89), Verbal Memory (0.81), Total Symptom scores (0.83), Reaction Time (0.77), and Visual Memory (0.63). The

Table 1. One-month test–retest reliability

Variable	Time 1	Time 2	*r*	ICC	ICC 95% CI		UER	*t*[a]	Sig	d[b]
					lower	upper				
Verbal Memory										
M	89.5	92.5	.66	0.788	0.518	0.906	0.805	−2.2	.04	0.50
SD	8.8	7.6								
Visual Memory										
M	69.6	71.4	.43	0.597	0.084	0.822	0.630	−0.8	.42	0.24
SD	11.2	9.8								
Visual Motor Speed										
M	39.0	41.8	.78	0.876	0.719	0.945	0.887	−3.6	.001	0.73
SD	5.8	6.0								
Reaction Time										
M	0.61	0.59	.63	0.767	0.471	0.897	0.786	0.8	.40	0.22
SD	0.08	0.06								
Symptom Scale										
M	7.4	9.1	.75	0.810	0.569	0.916	0.826	−0.9	.39	0.24
SD	9.1	14.6								

ICC = intraclass correlation coefficient;
UER = Unbiased Estimate of Reliability.
[a]$df = 24$; Bonferonni corrected alpha $p < .01$.
[b]Cohen's *d*, a measure of effect size.

Table 2. Reliable Change Indices

Variable	*r*	SE[1]	SE[2]	Sdiff	RCI	Δ80%[a]	Δ95%[a]
Verbal Memory	.67	5.15	4.47	6.82	−0.45	12 (88%)	14 (92%)
Visual Memory	.43	8.48	7.43	11.27	−0.16	19 (92%)	24 (96%)
Visual Motor Speed	.78	2.69	2.82	3.91	−0.73	6.3 (80%)	7.8 (84%)
Reaction Time	.63	0.047	0.039	0.061	0.17	0.10 (92%)	0.12 (96%)
Symptom Scale	.75	4.63	7.28	8.62	−0.20	15 (92%)	18 (92%)

r = Pearson's correlation between Time 1 and 2 scores; SE = Standard Error of Measure at Times 1 and 2 = $(SD \times \sqrt{(1 - r_{xy})})$; Sdiff = standard error of difference scores based on (Chelune et al., 1993–upper): $\sqrt{((SEM1^2) + (SEM2^2))}$; RCI = Reliable Change Index.
[a]Δ80% = absolute point change required for Reliable Change at 80% and 95% range (with the percent of participants with change scores within cut-off in parentheses).

mean ImPACT composite and symptom scores showed a significant improvement between the two assessments on only Motor Processing Speed (Table 1).

RCIs were calculated for all Composite and Total Symptom scores. RCIs are presented at 80% and 95% confidence intervals, in accordance with Chelune and colleagues (1993; Table 2). After a time interval of ~1 month, only a small percentage of participants' scores fell outside the range of normality, as denoted by these techniques. The RBM (Table 3) served as a more conservative measure, with no follow-up baseline scores falling outside of expected ranges. Regression-based measures, using 80% and 95% confidence intervals, revealed that follow-up baseline scores showed considerable stability (Table 3). All scores from follow-up baseline assessments fell within an 80% confidence interval; 88–91% of all follow-up baseline Composite and 86% of the follow-up Symptom Scale scores. Nearly all scores from follow-up baseline assessments fell within a 95% confidence interval; only a small percentage of scores from Visual Memory (1.0%), Reaction Time (0.2%), and Symptom Scores (2.4%) fell outside of the cut-off.

Given that the test–retest reliability of ImPACT has also been studied across 7 days, 45 days, 1 year, and 2 years, we compared the current data with Pearson's correlation coefficients and ICCs from published literature (where available, *see* Table 4). ICCs and/or Pearson's *r* all show a progressive pattern of decline from 7 to 30 days to 1 to 2 years, with two exceptions: (i) The 45-day test–retest reliability data from Broglio et al., (2007) is lower than data from all other time intervals and (ii) the Visual Memory scores from the current study are lower than data from 7 days, 1 year, and 2 years.

Discussion

This study examined the short-term test–retest reliability of assessments collected using the online version of the ImPACT test battery. The results suggest that college students may benefit from repeated exposures across a 1-month time period, but only in the

Table 3. Regression-based methods

Variable	Time 1	Time 2	α[a]	β[b]	Sxy[c]	80% CI[d]	95% CI[d]
Verbal Memory							
M	89.5	92.5	41.61	0.569	5.747	88%	96%
SD	8.9	7.6					
Visual Memory							
M	69.6	71.4	45.28	0.376	8.882	92%	100%
SD	11.2	9.8					
Visual Motor Speed							
M	38.9	41.8	9.86	0.820	3.774	88%	96%
SD	5.8	6.0					
Reaction Time							
M	0.60	0.59	0.276	0.525	0.049	88%	96%
SD	0.08	0.06					
Symptom Scale							
M	7.4	9.1	0.381	1.181	9.629	80%	96%
SD	9.3	14.6					

[a]α (alpha) = intercept.
[b]β (beta) = slope.
[c]Sxy = standard error of estimate (Crawford & Garthwaite, 2006).
[d]CI = Confidence interval; numbers represent the percent of participants with change scores within cut-off (80% CI = 1.65; 90% CI = 1.96).

Table 4. Comparison of Pearson's correlation coefficient and ICCs from published studies

Variable	Interval Between Assessments				
	7 days[a]	30 days[b]	45 days[c]	1 year[d]	2 years[e]
Verbal Memory					
ICC	—	.79	.23	.62	.46
r	.70	.66	—	.45	.30
Visual Memory					
ICC	—	.60	.32	.70	.65
r	.67	.43	—	.55	.49
Visual Motor Speed					
ICC	—	.88	.38	.82	.74
r	.86	.78	—	.74	.60
Reaction Time					
ICC	—	.77	.39	.71	.68
r	.79	.63	—	.62	.52

[a]Iverson et al., (2003); $n = 56$.
[b]Present study; $n = 25$.
[c]Broglio et al., (2007); $n = 73$.
[d]Elbin, Schatz, and Covassin (2011); $n = 369$.
[e]Schatz (2010); $n = 95$.

form of significant improvements in Visual Motor Speed scores. However, the remainder of the baseline composite scores remained relatively stable across a 1-month time period. Statistically significant between-assessment changes in Visual Motor Speed scores were supported by RCI data, showing a significant number of participants fell outside 80% and 95% confidence intervals. In contrast, all scores using the RBM were within 80% and 95% confidence intervals. Overall, these results suggest that repeated exposure to the ImPACT test across a 1-month interval does not result in practice effects in memory performance or reaction time. Comparison with ImPACT test–retest data in the literature reveals a steady pattern of decline, as a function of the time between assessments, with a few exceptions. Most notably, Broglio et al.,' (2007) Pearson's *r* and ICCs data are lower. In this regard, the present sample yielded only 3.8% ($n = 1$) invalid baselines, the time interval was only 2 weeks shorter (e.g., than their 45-day interval), and the current results are more in register with 1-year test–retest data. As such, questions about the methodological rigor of Broglio et al., (2007) in the context of completing four complete computer-based test batteries, on three occasions, comprising similar tasks and similar constructs may have contributed to variability in performance in their study. In addition, the online version of the ImPACT test uses the keyboard input for choice reaction time (e.g., different hands), whereas the desktop

versions uses mouse input (e.g., different fingers on the same hand). Although this revision has resulted in decreased "left-right confusion" and a subsequent decrease in associated elevations on Impulse Control scores (Schatz, Moser, Solomon, Ott, & Karpf, 2012), previous test–retest reliability research has utilized both the desktop version (Broglio et al., 2007; Iverson et al., 2003; Schatz, 2009) as well as the online version (Elbin et al., 2011), which may have contributed to variation in findings.

It is important to view these results in the context of its limitations. Participants were college students, but not varsity athletes, and may have presented to testing with significantly different motivations and interest. In addition, this was conducted as a pilot study, so the sample size is quite small, and not balanced by gender. In spite of these limitations, the current results were remarkably similar to previously reported test–retest results from larger samples (as cited in Table 4). Future research should focus on replicating these results: (a) in a larger sample, (b) in similar and longer time frames, (c) in athletes, and (d) across younger and older participants.

References

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

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.

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

Chicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6 (4)*, 294–290.

Crawford, J. R., & Garthwaite, P. H. (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: A significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology*, *20 (3)*, 259–271.

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

Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement*, *47*, 925–939.

Iverson, G. L., Lovell, M. R., & Collins, M. W. (2003). Interpreting change on ImPACT following sport concussion. *The Clinical Neuropsychologist*, *17 (4)*, 460–467. doi:10.1076/clin.17.4.460.27934.

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.

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

Lovell, M. R. (2007). *ImPACT Version 6.0 Clinical Interpretation Manual*. ImPACT Applications, Inc., Pittsburgh, PA.

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. doi:10.1080/13803395.2011.630655.

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.

Randolph, C. (2011). Baseline neuropsychological testing in managing sport-related concussion: Does it modify risk? *Current Sports Medicine Reports*, *10 (1)*, 21–26. doi:10.1249/JSR.0b013e318207831d00149619-201101000-00009 [pii].

Randolph, C., Lovell, M., & Laker, S. R. (2011). Neuropsychological testing point/counterpoint. *Physical Medicine and Rehabilitation*, *3 (10 Suppl. 2)*, S433–S439. doi:10.1016/j.pmrj.2011.08.002.

Randolph, C., McCrea, M., & Barr, W. (2006). Letters to the Editor. *Journal of Athletic Training*, *41 (2)*, 137–140.

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.

Rodgers, J. L., & Nicewander, W. A. (1998). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*, 59–66.

Schatz, P. (2010). Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *American Journal of Sports Medicine*, *38 (1)*, 47–53. doi:0363546509343805 [pii]10.1177/0363546509343805.

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. doi:10.1080/13803395.2012.667789.

Schatz, P., Moser, R. S., Solomon, G. S., Ott, S. D., & Karpf, R. (2012). Incidence of invalid computerized baseline neurocognitive test results in high school and college students. *Journal of Athletic Training*, *47 (3)*, 289–296. doi:10.4085/1062-6050-47.3.14.

Schatz, P., Neidzwski, K., Moser, R. S., & Karpf, R. (2010). Relationship between subjective test feedback provided by high-school athletes during computer-based assessment of baseline cognitive functioning and self-reported symptoms. *Archives of Clinical Neuropsychology*, *25 (4)*, 285–292. doi:acq022 [pii]10.1093/arclin/acq022.

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.

Solomon, G. S., & Haase, R. F. (2008). Biopsychosocial characteristics and neurocognitive test performance in National Football League players: An initial assessment. *Archives of Clinical Neuropsychology*, *23 (5)*, 563–577. doi:10.1016/j.acn.2008.05.008.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19 (1)*, 231–240. doi:10.1519/15184.1.

Wilk, C. M., Gold, J. M., Bartko, J. J., Dickerson, F., Fenton, W. S., Knable, M., et al. (2002). Test-retest stability of the Repeatable Battery for the Assessment of Neuropsychological Status in schizophrenia. *American Journal of Psychiatry*, *159 (5)*, 838–844.