2016

Test–retest, retest, and retest: Growth curve models of repeat testing with Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT)

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

Caitlin J. Masterson
University of Nebraska-Lincoln, cmasterson2@unl.edu

Tiffany D. James
University of Nebraska–Lincoln, tjames6@unl.edu

Jonathan Beckwith
Simbex LLC, Lebanon, NH

Per Gunner Brolinson
Virginia College of Osteopathic Medicine, Blacksburg, pbrolinson@vcom.vt.edu

See next page for additional authors

Authors
Arthur C. Maerlender, Caitlin J. Masterson, Tiffany D. James, Jonathan Beckwith, Per Gunner Brolinson, Joe Crisco, Stefan Duma, Laura A. Flashman, Rick Greenwald, Steven Rowson, Beth Wilcox, and Tom W. McAllister

Published in *Journal of Clinical and Experimental Neuropsychology,* 2016, 6pp; doi 10.1080/13803395.2016.1168781

Submitted January 4, 2016; accepted March 17, 2016.

Test–retest, retest, and retest: Growth curve models of repeat testing with Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT)

Arthur C. Maerlender,[1] Caitlin J. Masterson,[1] Tiffany D. James,[2] Jonathan Beckwith,[3] Per Gunner Brolinson,[4] Joe Crisco,[5] Stefan Duma,[6] Laura A. Flashman,[7] Rick Greenwald,[3] Steven Rowson,[5] Beth Wilcox,[4] and Tom W. McAllister[8]

1 Center for Brain, Biology and Behavior, University of Nebraska–Lincoln, Lincoln, NE
2 Office of Research and Economic Development, University of Nebraska–Lincoln, Lincoln, NE
3 Simbex LLC, Lebanon, NH
4 Department of Sports Medicine, Edward Via Virginia College of Osteopathic Medicine, Blacksburg, VA
5 Department of Bioengineering, Brown University, Providence, RI
6 Department of Bioengineering, Virginia Polytechnic Institute and State University, Blacksburg, VA
7 Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH
8 Department of Psychiatry, Indiana University School of Medicine, Indianapolis, IN

Corresponding author — Arthur C. Maerlender amaerlender2@unl.edu Center for Brain, Biology and Behavior, 1100 T Street, C89 East Stadium, University of Nebraska–Lincoln, Lincoln, NE, 68588-0156, USA.

Abstract
Computerized neuropsychological testing has become an important tool in the identification and management of sports-related concussions; however, the psychometric effect of repeat testing has not been studied extensively beyond test–retest statistics. The current study analyzed data from Division I collegiate athletes who completed Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) baseline assessments at four sequential time points that varied over the course of their athletic careers. Administrations were part of a larger National Institutes of Health (NIH) study. Growth curve modeling showed that the two memory composite scores increased significantly with successive administrations: Change in Verbal Memory was best represented with a quadratic model, while a linear model best fit Visual Memory. Visual Motor Speed and Reaction Time composites showed no significant linear or quadratic growth. The results demonstrate the effect of repeated test administrations for memory composite scores, while speed composites were not significantly impacted by repeat testing. Acceptable test–retest reliability was demonstrated for all four composites as well.

Keywords: Concussion, Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT), Growth curve modeling, Neuropsychology, Test–retest reliability

The interest and concern over sports-related concussions have led to procedures and protocols for identification and management of the injury. Neuropsychological testing has become an integral tool in the management of sports-related concussions and is recommended as part of an overall strategy (McRory, Meeuwisse, & Johnston, 2009). To better understand post-injury test results, individual baseline testing has become a recommended and integral part of many protocols (McRory et al., 2009). Comparing post-injury neuropsychological test scores to an athlete's own baseline is thought to provide a more objective and accurate measure of an individual athlete's performance. Thus, the effect of repeat testing is an important psychometric consideration in using these tests.

The Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) test is among the most widely used instruments for the assessment of sports-related concussion (Meehan, d'Hemecourt, Collins, & Comstock, 2012), and its psychometric properties have been extensively evaluated (Institute of Medicine, IOM, 2013). Mixed results have been obtained with regard to repeat test administrations (for example, see Elbin, Schatz, & Covassin, 2011; Nakayama, Covassin, Schatz, Nogle, & Kovan, 2014; Register-Mihalik et al., 2012; Schatz, 2010), suggesting both strong and weak interrater and test–retest reliabilities. Clearly, understanding the effect of repeat test administrations is an important consideration when monitoring effects of treatment or the course of recovery from injury. In fact, the American Academy of Clinical Neuropsychology (AACN) has recommended the use of statistical procedures for accounting for these effects when interpreting test scores (Heilbronner et al., 2010). In a study of relatively similar intervals to those in the present study, Schatz (2010) analyzed the test–retest reliabilities of ImPACT composite scores in 95 collegiate athletes after a 2-year interval. Improvements in composite scores were again modest, with intraclass correlations generally above .60, except for Verbal Memory (Verbal Memory = .459, Visual Memory = .642, Visual Motor Speed = .742, Reaction Time = .676).

A few studies have looked specifically at practice effects of ImPACT. Register-Mihalik et al. (2012) assessed age and practice effects of ImPACT in a mixed cohort of high-school and college athletes. Sample sizes were quite small (n = 20 in each cohort), and they obtained significant main effects for administrations (three) and age-group. The Visual Motor Speed composite was significantly different in the collegiate group between each time point, with the difference between Times 1 and 2 showing the greatest improvement. Between Test Administrations 1 to 2 and 2 and 3, 35% (n = 7) changed significantly according to reliable change indices. Over all, this composite had the fewest cases of significant change. Schatz and Ferris (2013) found significant improvements in the Visual Motor Speed composite score in a 25-day retest sample of 25 undergraduates. The absolute difference was 2.8 points with standard deviations greater than 5.8 points. In a much larger sample, Nakayama et al. (2014) noted small improvements in ImPACT composites across composite retest scores. Using regression-based change statistics and a 95% confidence interval (z-score change of 1.96), no Visual Motor Speed scores improved, and two declined across three test sessions. Verbal and Visual Memory composites had one improvement and five declines each, while Reaction Time had five improvements and one decline.

In a larger study of collegiate athletes, our study team had occasion to obtain baseline and noninjured ImPACT test results from multiple assessments over the course of four years (McAllister et al., 2012). With these data available, it became possible to address the question of the practice effect of multiple administrations of the ImPACT test in noninjured athletes. The present study sought to determine the growth trajectory of the four ImPACT composite scores across repeated test administrations, while allowing time to vary by individual.

Method

Overview

As part of our National Institutes of Health (NIH) study of the biomechanical basis of concussion and the effects of subthreshold impacts (grant number 1R01HD48638), athletes from three National Collegiate Athletic Association (NCAA) athletic programs underwent cognitive assessment (including ImPACT) at multiple time points: prior to the season, at post-season, and during the season (in some cases) over the course of four years (2008–2012). These noncontact athletes and contact sport athletes who were not concussed served as participants for various aspects of those studies. Thus, a large cohort of nonconcussed athletes taking the test over multiple administrations at various time points was obtained.

The original dataset had 333 collegiate athletes from three different Division I schools as part of a larger grant. The initial data cleaning identified nine cases that violated the ImPACT validity criteria on at least one test administration. We also eliminated four that had an improbable score of less than 5 on the Three-Letters Average Counted Correctly score (artificially improving the Verbal Memory composite score) and three that had scores of zero on the Color Match test. From the remaining 317 cases, we removed data for anyone who had sustained a concussion before or during the testing period (86), leaving a total of 231 participants. There were 161 contact and 70 noncontact participants at Test 1. The *contact sport cohort* consisted of football players at the three institutions and ice hockey players (men and women) from two of the three institutions. The *noncontact sport* cohort consisted of varsity athletes on a variety of teams (i.e., swim, cross-country, crew, track, golf, and softball). Athletes were excluded if they had significant systemic medical illness or current psychiatric disorders. All athletes were tested either as part of their standard athletic department procedures (i.e., supervised group testing), or as part of research procedures (i.e., individually administered by a trained technician). Overall, this sample was predominantly male, and the average age was 19.9 years.

Table 1 presents the numbers per test administration. The protocol was approved by the institutional review board (IRB) at each institution, and participants provided written informed consent.

Table 1. Number of participants, median number of days since first test, and composite means of each factor.

Test	N	Median days	Verbal memory	Visual memory	Visual motor	Reaction time
T1	231 (202, 29)	—	87.59 (8.64)	78.68 (12.64)	41.99 (5.88)	0.59 (0.12)
T2	231 (202, 29)	154	89.91 (9.57)	80.87 (11.63)	42.67 (6.25)	0.59 (0.12)
T3	127 (107, 20)	368	91.01 (8.40)	82.19(10.44)	42.57 (6.94)	0.59 (0.12)
T4	90 (72, 18)	488	91.59 (8.43)	82.14 (11.73)	42.41 (6.33)	0.58 (0.13)

Participants: male, female shown in parentheses. Standard deviations in parentheses

Materials

As part of the study research protocol, all participants completed ImPACT testing (Versions 2.0 and 2.1) at the beginning of each season and the end of the season. A mix of "baseline" and "post-injury" test forms were administered in no systematic manner. Most of the participants were administered the baseline form (Form 1) on ImPACT; however, an unknown number received Forms 2–4 at different administrations. ImPACT provides four composite test scores: Verbal Memory, Visual Memory, Visual–Motor Speed, and Reaction Time, as well as a total symptom scale score.

Analyses

Growth curve analysis was used to identify the score stability and practice effects over multiple administrations of the ImPACT test. All models were estimated using maximum likelihood estimation and were run with Mplus 7.31. Scores from the four ImPACT composites were examined in separate models: Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time. Reaction Time composite scores were log transformed to account for non-normal distribution.

The amount of time between tests was highly variable between participants, as well as the time intervals between each test per participant. To account for this variability in time between tests, individually varying time intervals were used to allow the time scores per participant to be the exact number of days since the baseline test. Our models allowed us to identify each participant's time interval between tests to more accurately examine practice effects (Muthén & Muthén, 2012). Estimating individually varying times of observation requires a random-effects-model, and typical fit indices (e.g., root mean square error of approximation, RMSEA; comparative fit index, CFI, etc.) and standardized estimates are not available (Muthén & Muthén, 2012).

Test–retest reliability was assessed with a two-way mixed-effects intraclass correlation (ICC) for the average of the four tests (absolute agreement).

The independent variables included site of administration, sex, the presence or absence of attention and/or learning problems, and total current symptoms at the time of testing.

Results

The means and standard deviations of each composite score generally show small increases per administration for the two memory composites (verbal, visual), while the speed composites appear to remain relatively invariant (Visual Motor Speed, Reaction Time: Table 1). Covariate analyses determined that site of administration, sex, special education, and symptoms were not significant predictors of outcomes. These time-invariant variables were analyzed as part of the growth curve models and were found to have no significant effect.

Test–retest reliability (ICC) was significant for all four composites at $p < .001$: Verbal Memory = .737, Visual Memory = .776, Visual Motor Speed = .893, Reaction Time = .773 ($N = 126$).

Growth curve analysis on the Verbal Memory composite factor revealed a significantly positive quadratic term as the highest polynomial necessary to accurately describe practice effects over multiple test administrations (Table 2). The growth curve for this score demonstrated the greatest improvement from Test 3 to Test 4, indicating a nonlinear increase in scores over test administrations. Figure 1 shows the expected values for Verbal and Visual Memory composite scores at each assessment using the intercept, linear (quadratic) value, and time. The Visual Memory composite factor model revealed a significant, positive linear growth (Table 2).

Table 2. Estimates and *p*-values of the linear and quadratic models of each factor.

	Linear		Quadratic	
Composite	Estimate	*p*	Estimate	*p*
Verbal memory	0.00371	.008	**0.00017**	**.015**
Visual memory	**0.00343**	**.042**	0.00009	.238
Visual motor	–0.00013	.881	0.00010	.082
Reaction time	0.00001	.860	0.00002	.275

Significant estimates shown in bold.

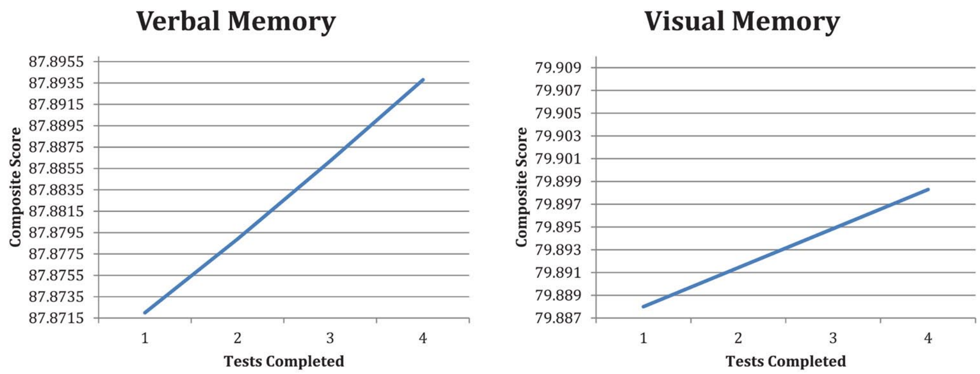

Figure 1. Growth curves for Verbal and Visual Memory composites.

Thus, over time, with each test administration, Visual Memory composite scores increased at a constant rate. Both the Visual Motor Speed and Reaction Time composite scores demonstrated no significant linear or quadratic changes over time. These results demonstrate that practice effects for these factors are minor, when controlling for the length of time between test administrations. Note that the growth curves in Figure 1 are scaled in order to demonstrate the curves, although the absolute changes are quite small (but statistically significant).

Discussion

This study examined the effect of multiple exposures to the computerized ImPACT test in a cohort of college athletes who were not diagnosed with concussion. By using growth curve modeling we were able to examine the change curves as a function of test exposure while controlling for precise time intervals between test administrations. The Verbal Memory composite scores increased significantly over the four test administrations, defined by a quadratic function: Scores improved at each time point, but with increased growth from Test 3 to Test 4. The Visual Memory composite scores fit a linear model, indicating that scores improved at the same rate at each successive testing. Visual Motor Speed and Reaction Time did not improve significantly across the four time points and fit no model. Although statistically significant, the amount of score change in memory scores was actually quite small on average. This obscures the importance of the absolute level of each score, which impacts change calculations greatly.

Although not the focus of this study, our findings are generally consistent with other test–retest studies of ImPACT. In three studies that compared repeat testing score changes for reliability purposes, the actual amount of score change was quite small for Visual Motor Speed and Reaction Time, and greater for the two memory composites (Nakayama, et al., 2014; Register-Mihalik et al., 2012; Schatz, 2010; Schatz & Ferris, 2013). The maximum score change for Visual Motor Speed was 2.8 points, and for Reaction Time the maximum score difference was .02, both from Schatz and Ferris (2013). Other test–retest papers looked at mixed samples or were not comparable due to methodological limitations (e.g., Broglio, Macciocchi, & Ferrara, 2007).

The clinical significance of these results is that practice effects were significant for the memory composites (Verbal, Visual), but were limited in the two speed composite scores (Visual Motor Speed and Reaction Time). The data indicated that Visual Memory scores increased consistently at subsequent test administrations, and Verbal Memory scores show acceleration in change between Tests 3 and 4.

Test–retest reliability was stronger than that in some studies, likely due to the average across the four measures and the large sample size, together with rigorous case selection. Of the reliability studies explored, only Schatz (2010) used intervals approaching the length of time here (2 years). In general, the longer intervals may have a positive effect on reliability as well, although it is not clear why it would be so. However, the pattern was consistent with that in previous studies showing Visual Motor Speed with the strongest reliability, and Verbal Memory with the weakest.

The differentiation between practice effects on memory and speed composite scores is theoretically interesting. These composite score combinations have been shown to be factorially related and demonstrated improved reliability over the individual composite test scores (Schatz & Maerlender, 2013).

The findings extend our understanding of score changes due to repeat testing. They suggest that the memory composite scores can be expected to show practice effects out to at least four test administrations. The speed composite scores are more stable and did not change significantly over multiple administrations.

As with any study, several limitations need to be acknowledged that limit interpretation and generalizability. First, this is a sample of college athletes, and therefore the age range may limit inference to high-school-aged and younger students, or older adults. There was also limited control of test conditions as most of these results came from group administration procedures that have been shown to reduce reliability (Moser, Schatz, & Lichtenstein, 2013; Moser, Schatz, Neidzwski, & Ott, 2011). The relatively small sample of females may also bias the sample, and no analysis by race or ethnicity was undertaken due to missing demographic data. Certainly, replication with other samples would be highly desirable.

An important consideration involves the clinical use of alternative forms for testing. Most variance in retesting is felt to be due to familiarity with the procedure as opposed to the specific content (Lezak, Howieson, Bigler, & Tranel, 2012). One study has attempted to compare the four alternate forms of ImPACT (Resch, Macciocchi, & Ferrara, 2013). They used inferential confidence intervals to establish the equivalency of the various test forms. They found that, on the whole, the Visual Motor Speed and Reaction Time alternate forms were equivalent, while Verbal Memory and Visual Memory were quite variable. They did not control for the practice effects or regression to the mean due to repeat testing so it is difficult to take much from their findings as this would be a critical factor for interpreting any retest analysis. However, the point is well taken that some variance is introduced by having different content. This is not considered in the ImPACT manual or in ImPACT's calculation of reliable change intervals. Thus while theoretically relevant, this should not render this approach to change scores useless.

This is the largest known sample of multiple repeat test administrations with the ImPACT test and as such provides useful data for understanding the stability of composite scores over time. Further studies with other populations would be valuable.

Disclosure — No authors have any financial interest or have benefited from the direct application of this research.

Funding — This work was supported by National Institute of Health National Center for Medical Rehabilitation Research [grant number BRP R01HD048638]; National Institute of Child Health and Development; and National Institute of Health/National Institute of Neurological Disease and Stroke [grant number RO1NS055020].

ORCID — Arthur C. Maerlender 0000-0001-9740-6354

References

Broglio, S. P., Macciocchi, S. N., & Ferrara, M. S. (2007). Sensitivity of the Concussion Assessment Battery. *Neurosurgery*, *60*(6), 1050–1058.

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*, *39* (11), 2319–2324.

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *Clinical Neuropsychologist*, *24*(8), 1267–1278.

Institute of Medicine (IOM) and National Research Council (NRC). (2013). *Sports-related concussions in youth: Improving the science, changing the culture*. Washington, DC: The National Academies Press.

Lezak, M., Howieson, D. B., Bigler, E., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford.

McAllister, T. W., Flashman, L. A., Maerlender, A., Greenwald, R. M., Beckwith, J. G., Tosteson, T. D., . . . Turco, J. H. (2012). Cognitive effects of one season of head impacts in a cohort of collegiate contact sport athletes. *Neurology*, *78*(22), 1777–1784.

McRory, P., Meeuwisse, W., & Johnston, K. (2009). Consensus statement on concussion in sport: The 3rd International Conference on Concussion in Sport held in Zurich, November 2008. *The Physician Sportsmed*, *37*(2), 141.

Meehan, W. P., d'Hemecourt, P., Collins, C. L., & Comstock, R. D. (2012). Assessment and management of sport-related concussions in United States high schools. *The American Journal of Sports Medicine*, *39*(11), 2304–2310.

Moser, R., Schatz, P., & Lichtenstein, J. (2013). The importance of proper administration and interpretation of neuropsychological baseline and postconcussion computerized testing. *Applied Neuropsychology: Child*, *4*(1), 41–48.

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.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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*, *42*(8), 2000–2005.

Register-Mihalik, J. K., Kontos, D. L., Guskiewicz, K. M., Mihalik, J. P., Conder, R., & Shields, E. W. (2012). Age-related differences and reliability on computerized and paper-and-pencil neurocognitive assessment batteries. *Journal of Athletic Training*, *47*(3), 297–305.

Resch, J. E., Macciocchi, S., & Ferrara, M. S. (2013). Preliminary evidence of equivalence of alternate forms of the ImPACT. *Clinical Neuropsychology*, *27* (8), 1265–1280.

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

Schatz, P., & Ferris, C. S. (2013). One-month test–retest reliability of the ImPACT test battery. *Archives of Clinical Neuropsychology*, *28*(5), 499–504.

Schatz, P., & Maerlender, A. (2013). A two-factor theory for concussion assessment using ImPACT: Memory and speed. *Archives of Clinical Neuropsychology*, *28*, 791–797.