Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues

Lester B. Mayers¹ and Thomas S. Redick²

¹Division of Sports Medicine, Pace University, Pleasantville, NY, USA
²Division of Science, Indiana University-Purdue University Columbus, Columbus, IN, USA

Computerized neuropsychological testing is commonly utilized in the management of sport-related concussion. In particular, the Immediate Postconcussion Assessment and Cognitive Testing 2.0 program (ImPACT) is widely used to assess the cognitive functioning of athletes before and after a concussion. We review the evidence for the clinical utility of this program in terms of validity, reliability, and use in return-to-play decisions. We conclude that the empirical evidence does not support the use of ImPACT testing for determining the time of postconcussion return to play.

Keywords: Athlete; Concussion; Neuropsychological testing; Reliability; Validity.

Our current evaluation techniques for a patient suffering from concussion are inadequate. . . . The best new technique for evaluating concussion is cognitive and psychometric testing that begins with a preseason screening that can be used as a baseline when compared with congruent tests performed after concussion occurs. Return to baseline on a psychometric test can help determine when a player can safely resume strenuous exercise and body contact. (Ashare, 2009, p. 775)

ImPACT . . . can also provide post-injury cognitive and symptom data that can assist a practitioner in making safe return to play decisions. (Schatz, Pardini, Lovell, Collins, & Podell, 2006, p. 91)

Our preliminary data suggest neurocognitive testing should be employed in all instances when making postconcussive return-to-play decisions. (Broglio, Macciocchi, & Ferrara, 2007, p. 507)

Computerized neuropsychological testing (CNT) for postconcussion assessment and return-to-play counseling is commonly utilized in sports medicine practice. There are many advantages of CNT over traditional pencil-and-paper tests, including ease of administration, testing with instant scoring for many athletes at one session, precision of response times, and minimal time demand for the test administrator (Brown, Guszkiewicz, & Bleiberg, 2007; Lovell, Collins, & Bradley, 2004). However, the basic premise underlying the use of neuropsychological tests for the clinical management of concussed athletes remains unclear. While the ability to efficiently assess neurocognitive status might, in theory, facilitate return-to-play decisions, clinical utility requires sound psychometric properties of the test instrument and validity of the assumptions underlying its use. Herein, we review evidence concerning Immediate Postconcussion Assessment and Cognitive Testing (ImPACT) testing pertaining to the quotes above and question the basic premise for applying this tool for the clinical management of concussed athletes. We do not, however, address the role of paper-and-pencil tests for this purpose (for review, see Randolph, McCrea, & Barr, 2005).

Among the several currently utilized CNT instruments are Automated Neuropsychological Assessment Metrics (ANAM), CogState Sport®, HeadMinder, and Immediate Postconcussion Assessment and Cognitive Testing (ImPACT). These have been comprehensively reviewed (Randolph et al., 2005), with the authors concluding that "additional research is clearly necessary.
before NP testing can be considered a component of the routine standard of care in the management of sport-related concussion, particularly as the risks of premature return to play remain poorly defined” (p. 145).

We focus our review on ImPACT because it is “the first, most widely used, and most scientifically validated computerized concussion evaluation system” (www.impacttest.com). Covassin, Elbin, Stiller-Ostrowski, and Kontos (2009) reported that 1,209 high schools and colleges currently use ImPACT. It is also utilized by many professional sports leagues around the world (www.impacttest.com). However, the assertion that ImPACT is a “reliable and valid tool in the assessment of concussion” (Majerske et al., 2008, p. 266) has been questioned (Kirkwood, Randolph, & Yeates, 2009; Randolph, 2011; Randolph et al., 2005). These authors have discussed many unresolved issues concerning ImPACT; therefore we highlight only the most relevant aspects here. We review the evidence for the clinical utility of ImPACT in terms of validity, reliability, and use in return-to-play decisions. ImPACT contains three main sections: (a) demographic and health history information; (b) a 22-item current symptom inventory; and (c) six test modules that are combined to create composite scores representing processing speed, reaction time, visual memory, and verbal memory. Readers interested in more information about ImPACT should refer to the website of the company that sells the test (www.impacttest.com).

Ideally, ImPACT would provide accurate measures to answer two important questions:

1. Should the athlete participant limit and/or forgo further sport-related exposure to head injury in order to avoid the potential for future chronic disability?
2. When is it safe to approve an athlete’s return to play after a recent concussion?

All published studies relating to the first question indicate that ImPACT cognitive performance normalizes within 3 weeks post injury. Investigators have reported subtle, lasting visuomotor processing impairments among head-contact-exposed rugby players with the pencil-and-paper Digit Symbol and Trail Making tests (Shuttleworth-Edwards & Radloff, 2008), and residual attention (Moser & Schatz, 2002) and memory (Killam, Cautin, & Santucci, 2005) impairments with the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) in youth and collegiate athletes with prior concussions. In contrast, several groups utilizing ImPACT found normal results among previously concussed but recovered athletes (Broglio, Ferrara, Piland, Anderson, & Collie, 2006; Iverson, Brooks, Lovell, & Collins, 2006; Solomon & Haase, 2008). The same finding using a computer-based test of working memory capacity was reported recently (Mayers, Redick, Chifferiller, Simone, & Terraforte, 2011). Therefore, current computerized test instruments do not provide early alerts that might prevent the subsequent development of chronic traumatic encephalopathy (Omalu et al., 2006) among concussed and multiconcussed athlete cohorts. We also note that ImPACT is not marketed as a tool for assessing long-term impairments related to previous sports-related brain injury. However, ImPACT was an outcome measure used in a study designed to “determine the long-term effects on the brain of a career in the National Football League” [NFL] (Schwarz, 2009, p. B10). We note this particular study because two developers of ImPACT (Maroon and Lovell) were part of the committee overseeing the study for the NFL. Although postconcussion cognitive recovery appears to occur relatively quickly, this does not equate to rapid healing of the injured brain—an issue that we address below (see Discussion).

The answer to the second question (viz., assisting return-to-play decisions) is controversial, but herein we review published findings and present our conclusions. The proper time interval between the occurrence of a brain injury (concussion) and the time of safe return to play is a medical decision and should ideally be evidence based. Tests utilized to achieve this purpose should be reliable (attaining acceptable test–retest repeatability), sensitive (reliably detecting an affected patient), and specific (not falsely identifying normal subjects as being affected), and have acceptable validity (sensitive to impairment when true impairment is present).

**TEST–RETEST RELIABILITY**

A significant concern when considering a role for CNT in concussion management involves the reliability of the measurements. There are many different types of reliability, but most germane for the use of CNT such as ImPACT is test–retest reliability, which is the type we focus on (in contrast to, for example, internal consistency of test items). We note that there is no single measure of test–retest reliability—as noted by Anastasi and Urbina (1997), “since retest correlations decrease progressively as this interval lengthens, there is not one but an infinite number of retest reliability coefficients for any test” (p. 92). Obviously, a highly reliable
test is preferred to a less reliable one. However, there are always practical considerations (e.g., time of administration, cost of administration) to consider when weighing the clinical utility of tests as well. Murphy and Davidshofer (1998) note that high test reliability is most important when tests are used to make assignments or decisions about people, similar to what occurs in the baseline-testing application of CNT (impaired versus not impaired). Quoting Murphy and Davidshofer (1998, p. 142), “For most testing applications, reliability estimates around .70 are likely to be regarded as low. . . reliability estimates lower than .60 usually are thought to indicate unacceptably low levels of reliability.” In view of this information, we review the three published studies that have assessed test–retest reliability in ImPACT.

Iverson, Lovell, and Collins (2003) reported Pearson $r$ test–retest reliabilities between .65 and .86 in a sample of 56 healthy controls tested 1–13 days apart. Importantly, the authors noted that the interval used was very short compared to the clinical application of ImPACT. That is because ImPACT’s purveyors recommend performing baseline tests twice during high school and once in college (www.impacttest.com/faq#faq_11). Therefore the interval between baseline and postconcussion testing will likely be months or years rather than days. Two recent studies measured the test–retest reliability over longer intervals (Broglio, Ferrara, Macciocchi, Baumgartner, & Elliott, 2007; Schatz, 2010). Broglio, Ferrara, et al. (2007) reported test–retest intraclass correlations of .23 to .38 for a sample of 76 healthy controls tested 45 days apart. Schatz (2010) reported test–retest intraclass correlations of .45 to .74 and Pearson $r$ test–retest correlations of .30 to .60 in a sample of 95 athletes tested approximately 2 years apart. Based on the Murphy and Davidshofer (1998) criteria provided above, the test–retest correlations reported by Iverson, Lovell, & Collins, (2011) tested twice within 2 weeks would be acceptable, but the ImPACT reliabilities reported in the studies (Broglio, Ferrara, et al., 2007; Schatz, 2010) using more clinically relevant test intervals, and larger samples would be considered “unacceptably low.”

Lack of reliability is a major concern for any test but especially so because ImPACT users are encouraged to use reliable change confidence intervals to make return-to-play decisions about individual athletes (Iverson, Lovell, et al., 2003; Van Kampen, Lovell, Pardini, Collins, & Fu, 2006). Since the test–retest coefficient is an integral part of the formula for assessing statistically significant change, using the values provided by either Broglio, Ferrara, et al. (2007) or Schatz (2010) results in extremely large confidence intervals that would seriously limit interpretation in a clinical application. For example, based on the values reported in Schatz (2010), athletes tested on ImPACT after concussion would have to show a substantial decrease in verbal memory performance to be classified as significantly below their baseline test. This increases the likelihood that a true cognitive impairment is missed (decreased sensitivity). We also note that the reliable change confidence intervals provided by Iverson, Lovell, et al. (2003) are based on sample of 56 combined high-school and collegiate males and females. Given that the available norms have indicated gender and age differences in performance on ImPACT composites (Iverson, Lovell, & Collins 2011), reliable change confidence intervals would need to be computed from a much larger sample and provided separately for the appropriate ages and genders so as to have the most accurate identification of impairment.

Another issue affecting reliability is performance improvement during repeated testing (referred to as learning or practice effect). Although a purported advantage of ImPACT is that it “is not subject to the large practice effects sometimes seen on pencil and paper tests” (Schatz et al., 2006, p. 93), significant practice effects in healthy controls have been observed for the processing speed (Iverson, Lovell, et al., 2003) and visual memory (Schatz, 2010) composites. The presence of these practice effects will hamper interpretation of postconcussion testing, where return-to-play decisions are based on scores returning to baseline. This affects testing because practice effects that lead to improved performance may mask impairments that are decreasing performance. In addition, the calculation of the reliable change confidence intervals will have to be modified to account for the significant change in scores expected from practice.

VALIDITY

Randolph et al. (2005) critically reviewed the various aspects of validity evidence in CNT, but we focus upon the relevant studies related to return-to-play decisions. Most research to date has measured the sensitivity of ImPACT results to concussive injury. These studies have administered ImPACT to recently concussed athletes and compared their scores versus baseline, controls, or normative data. When recently concussed athletes are tested with ImPACT while they are still symptomatic, scores tend to be lower than those at baseline or for a noninjured control group. In this sense, ImPACT
does possess construct validity—the test is sensitive to the construct of “recent concussion history.” However, we agree with the consensus statement of the 2nd International Conference on Concussion in Sport (McCrory et al., 2005) that CNT performed during the concussed athlete’s symptomatic interval is not useful since he/she will not be cleared to play in any case.

Another important aspect of validity for ImPACT is the degree of divergent validity among the four composites assessed. Unfortunately, there is little peer-reviewed evidence available concerning this issue. In addition, the intercorrelations are not published as part of the normative data available for download on the ImPACT website (Iverson, Lovell, et al., 2003). A conference abstract published by Iverson, Lovell, and Collins (2002) notes that divergent validity was assessed via correlations among composites obtained during baseline testing of 120 high-school and college athletes. Although the abstract notes that “small correlations indicate that the composite scores do not have much shared variance” (Iverson et al., 2002, p. 769), the values are not reported. It also seems that Iverson et al. (2002) used ImPACT 1.0 before changes in the procedure and numbers of composites were made for the current version of ImPACT.

The only other study of healthy controls we located that reported intercorrelations for ImPACT composites was by Schatz and Putz (2006). In the context of assessing correlations among three different CNT instruments, they reported only the visual and verbal memory correlation ($r = .340$). Although their conclusion was that the composites were not significantly related, which might be taken as evidence for divergent validity, their sample was extremely small for correlational research ($N = 30$). In fact, the $p$-value for this nonsignificant correlation was .06. A correlation of $r = .340$ would be significant ($p < .05$) in a sample with only 4 more participants. Unfortunately, this was the only ImPACT intercorrelation reported by Schatz and Putz (2006). We argue that this lone, nearly significant correlation does not constitute strong evidence for divergent constructs.

Two other studies (Iverson, Franzen, Lovell, & Collins, 2003; Iverson, Lovell, & Collins, 2005) have addressed the divergent validity issue, but only within the context of athletes still suffering from the acute effects of concussion. Given this limitation, the following results should be taken with caution. Iverson, Franzen, et al. (2003) observed significant correlations ranging between .50 and .75 among the four ImPACT composites. In addition, although Iverson et al. (2005) do not report the actual correlation, they note in their study that the processing speed and reaction time composites measured “a similar underlying construct” (p. 687). The lack of separate constructs is not a trivial matter. First, if the composites share substantial overlap in variance, then an individual who is temporarily impaired in one ability (e.g., processing speed) will be likely to show impairment in a related ability (e.g., reaction time). If this is the case, interpreting the sensitivity of the ImPACT reliable change index scores across composites is weakened. However, exhibiting multiple composite scores outside of the confidence intervals suggested by Iverson, Lovell, et al. (2003) need not reflect severity of symptoms; rather, it could reflect the degree of intercorrelation among the composites.

If responsible sports medicine practitioners are influenced by the number of composites that have not returned to baseline in a recently concussed athlete, this may unnecessarily prolong their return to play. If the composites are orthogonal (viz., uncorrelated), then one may correctly interpret impairments in multiple cognitive domains. Based on the evidence reported above about nearly significant correlations between the verbal and visual memory composites (Schatz & Putz, 2006) and that the processing speed and reaction time composites measure a similar underlying construct (Iverson et al., 2005), it appears that the composites are not orthogonal. If the composites are indeed measuring similar abilities, the obvious question is why all four domains are being assessed. If the same or similar information can be gained by administering fewer tasks, this becomes an important consideration for clinical application.

**RETURN-TO-PLAY STUDIES**

To date, there are three reports that have addressed the use of ImPACT for return-to-play decisions via study of athletes who were administered ImPACT at baseline and then serially after suffering a concussion. There are a number of potentially important methodological differences among the following three studies in terms of: (a) the number of athletes tested; (b) the ages of the samples; and (c) the number and timing of postconcussion test sessions. These differences could affect the particular results of any of the studies. However, we focus on the information pertaining to return-to-play decisions.

McClincy, Lovell, Pardini, Collins, and Spore (2006) tested 104 athletes (88% male, 12% collegiate) approximately 2, 7, and 14 days post concussion. Self-report symptoms were still significantly elevated at Day 2 and Day 7, so we focus on the test results of Day 14. The verbal memory composite
was the only score significantly impaired relative to baseline at Day 14. While this study provides preliminary evidence that ImPACT is sensitive to impaired cognition while an athlete is symptomatic, we note that there was considerable variability associated with the 14-day test date ($M = 14.35, SD = 7.34$). Because of this variability, it is possible that some athletes underwent their third postconcussion test session around Day 7 or 8 when they were still symptomatic.

Iverson, Brooks, Collins, and Lovell (2006) administered ImPACT to 30 athletes (93% male, 10% collegiate) four times: baseline, and 1–2, 3–7, and 7–21 days post concussion. Again, self-report symptoms were still significantly elevated at the 1–2- and 3–7-day test periods, so we focus on the results of the 7–21-day test period. Critically, none of the ImPACT composites was significantly different from baseline. The authors noted a selection bias that “increased the probability that we would study athletes who were more seriously concussed, slow to recover, or both” (Iverson, Brooks, Collins, et al., 2006, p. 249). Despite this bias, they found no significant impairments when athletes were asymptomatic, in contrast to the results of McClincy et al. (2006).

Covassin, Elbin, and Nakayama (2010) conducted the most extensive prospective study of ImPACT. Seventy-two high-school athletes (76% male) completed ImPACT at baseline and exactly 2, 7, 14, 21, and 30 days post concussion. Self-report symptoms were higher at Day 2 ($p < .01$) and Day 7 ($p = .06$), so we focus on the results for Days 14, 21, and 30. There was one significant result out of 12 comparisons with baseline: Reaction time was significantly impaired at Day 14.

In summary, these three studies reported three different results using repeated ImPACT testing of concussed athletes. Clearly, the utility of ImPACT for clinical application can be questioned based on these inconsistent findings. In fact, the most consistent result across studies was that when athletes were no longer symptomatic by self-report, performance on the four ImPACT composites was typically not impaired relative to baseline. Another limitation of these studies is that the duration of symptoms for individual subjects was not reported. Therefore, it is not known how many participants were still symptomatic even at the longer postconcussion intervals.

**OTHER ISSUES**

The logic of the baseline approach to postinjury recovery is that a test should identify individuals that are: (a) performing worse than their baseline during the injury phase, and (b) performing nonsignificantly different than their baseline after recovery has occurred. Because multiple test administrations are inherent in this design, it is essential that the tests show high levels of test–retest reliability to be able to delineate significant and nonsignificant changes from baseline. As outlined by Campbell and Stanley (1963), there are a number of methodological considerations to consider when employing a pretest–posttest design. First, one needs a carefully matched control group that is tested over the same test–retest interval as the patient group. Without a control group, there is no way of determining whether significant change from baseline in the patient group is due to history, maturation, testing, instrumentation, selection, attrition, and regression-to-mean effects. Broglio, Ferrara, et al. (2007) suggested that subject “sleep deprivation and stimulant use, intense physical activity and daily stressors” (p. 513) might contribute to test–retest differences in athletes, which again indicates the need to use a carefully matched control group for return-to-play studies.

In addition, the design of ImPACT implies that multiple statistical comparisons will be needed—that is, comparisons between subjects (concussed versus control group), within-subjects (baseline versus the number of postconcussion test sessions), and across composites (four cognitive and one symptom). The use of multiple cognitive constructs in ImPACT is a strength, as opposed to creating one aggregate score from all of the modules administered. However, from a statistical standpoint, it suggests at the very least that given the lack of theory-driven reasons why one particular cognitive composite over the others may be impaired in concussed athletes, any analyses of ImPACT should begin with a multivariate analysis of variance. If the multivariate analysis is significant, then the appropriate follow-up univariate analyses can be conducted, with the necessary post hoc alpha level correction. Researchers should ensure that the appropriate correction for multiple statistical comparisons is used to control the familywise error rate in studies using ImPACT.

**DISCUSSION**

The National Athletic Trainer’s Association position paper (Guskiewicz et al., 2004) concerning management of sport-related concussion emphasized that CNT was not a diagnostic tool for concussion but might be valuable for measuring the time of recovery after a clinical diagnosis. They
stated that measuring recovery was “a complex statistical matter . . . complicated by practice effects and other psychometric dynamics affected by serial testing” and concluded that further research was required to determine guidelines pertaining to post-concussion recovery and appropriate interval for return to play “above and beyond the simple conclusion that the player is back to baseline” (Guskiewicz et al., 2004, pp. 289–290). The evidence we have reviewed here fits with this recommendation and indicates that CNT such as ImPACT does not meet the reliability and validity criteria desired in a test used for assisting return-to-play decisions.

We therefore question the rationale of using ImPACT for clinical management of sport-related concussion and specifically for determining the time of return to play. We should note that no “gold standard” exists for concussion diagnosis and management. Sports medicine practitioners still lack simple, reliable, and affordable techniques to confidently address these issues. Although experienced first responders can accomplish a clinical diagnosis in most instances of suspected concussion, concerns related to return to play and whether or not to continue specific sport participation are not resolved by current CNT.

Many studies using various measures of cerebral function in brain-injured individuals demonstrate impairment well beyond the interval within which most concussed athletes return to play (for review, see Mayers, 2008). These include gait stability (Parker, Osternig, Van Donkelaar, & Chou, 2006), balance studies (Slobounov, Slobounov, Sebastianelli, Cao, & Newell, 2007), evoked potentials (De Beaumont, Brisson, Lassonde, & Jolicoeur, 2007; De Beaumont, Lassonde, Leclerc, & Theoret, 2007), magnetic resonance spectroscopic measurements of brain-injury-associated metabolites (Vagnozzi et al., 2008), histological studies demonstrating diffuse axonal injury (Blumbergs et al., 1994), and diffusion tensor imaging (Cubon, Putukian, Boyer, & Dettwiler, 2011). All of these techniques have revealed abnormal function up to 28 days post injury and beyond. Although ImPACT may detect reduction from baseline scores in one or more neuropsychological domains after postconcussion symptoms have resolved, their clinical significance and the significance of their resolution are not established. We believe that measured functional and metabolic impairments are more clinically relevant than deficits observed in a single (or several) neuropsychological domain(s) because they persist for significantly longer time intervals and more specifically define functional neurological deficits. In a direct test of this idea, Broglio, Pontifex, O’Connor, and Hillman (2009) administered ImPACT to athletes and compared performance as a function of concussion history. They found no differences on any of the ImPACT composites, yet event-related brain potentials were significantly different between the concussed and nonconcussed groups when examined on a separate task.

In fact, the fundamental rationale for using CNT for postconcussion diagnosis and management depends on a hypothesis—namely, that baseline and postinjury tests are a reliable and objective measure of cerebral dysfunction and brain injury and that resolution of postconcussion neuropsychological impairment(s) equates to healing of the injured brain. In view of the many issues concerning reliability and validity we have reviewed, it seems that the current version of ImPACT fails to meet this standard of proof.

The attraction of CNT, as mentioned previously, is its relative ease of performance and availability in comparison with the clinical techniques cited above. It may therefore be argued that while routine CNT of concussed athletes may not offer much clinical benefit, it will not hurt. This will not be true, however, if the lay public (athletes, parents, coaches, and administrators) and medical professionals assume that the return of a neuropsychological test battery to “baseline” following concussion symptom resolution “proves” that the athlete is fit to play and that this assumption is allowed to influence clinical decisions. Furthermore, programs such as ImPACT have financial and logistical costs and can involve potential risk to athletes if they are required to travel to a testing site specifically to undergo the test. Finally, the only treatment currently available for concussion is physical and cognitive rest. We would not advise an athlete to take a final examination while he/she was acutely concussed. Requiring an athlete to undergo a 20-minute detailed neurocognitive examination (see ImPACT technical manual) in order to confirm a clinically apparent concussion is unnecessary and may be injurious.

Our conclusions are in agreement with a recently published cogent review that stated, “Despite the widespread use of baseline neuropsychological testing in the management of sport-related concussion, there appears to be essentially no evidence in the medical literature to suggest that this approach has modified any associated risks” (Randolph, 2011, pp. 25–26). In summary, for evaluating and advising concussed athletes when to return to play, ImPACT test results should not be the determining factor. Clearly, more research and consideration by neuropsychologists will be necessary to weigh the validity of this conclusion.
REFERENCES


