Psychometric issues associated with computerised neuropsychological assessment of concussed athletes.

Alexander Collie PhD, Paul Maruff PhD, Michael McStephen BSc(Hons) & Michael Makdissi MBBS

From 1. CogState Ltd
Carlton, Victoria, Australia

2. Centre for Sports Medicine Research and Education
The University of Melbourne
Parkville, Victoria, Australia

3. Centre for Neuroscience
The University of Melbourne
Parkville, Victoria, Australia

4. School of Psychology
La Trobe University
Bundoora, Victoria, Australia

Conflict of interest: Authors Collie, Maruff & McStephen are employees and equity holders in CogState Ltd, the company that develops and markets the CogSport computerised cognitive testing software.

Address for correspondence:
Dr Alex Collie
51 Leicester Street
Carlton 3053, Victoria
Australia
Phone: +613 9349 1300
Fax: +613 9348 2689
Email: acollie@cogstate.com
Introduction

The dual roles of neuropsychological testing in sports concussion are well established. Neuropsychological assessment may aid understanding of the brain structures and processes underlying concussion and the post-concussion syndrome. Although this is a primary goal of neuropsychologists working in sport concussion, a more immediate role lies in facilitating effective post-concussion medical management of individual athletes. In this context, neuropsychological tests may aid both detection of post-concussive cognitive impairments and provide a de facto measurement of brain function to assist return-to-play decision-making\(^1\). Over the past two decades, ‘paper-and-pencil’ neuropsychological tests have been used to aid the medical management of concussed, professional athletes in many sports\(^2\text{-}^5\). The psychometric and practical limitations associated with paper and pencil tests (reviewed in Collie et al\(^6\)) has led to the development of a number of computerised neuropsychological test batteries\(^7\text{-}^8\). This brief paper introduces some of the psychometric issues associated with computerised neuropsychological assessment in sports concussion. A number of issues critical to ensuring test reliability and sensitivity are discussed, with particular reference to how inappropriate test design can affect clinical decision making.

A driving factor behind the rapid adoption of computerised neuropsychological testing is the assumption that computerised tests are both more reliable and more sensitive to concussion related cognitive deficits than paper-and-pencil tests. The assumption of enhanced reliability appears to be based mainly on manufacturers’ claims that computerisation of neuropsychological tests reduces administrator bias, standardises task administration, allows randomised stimulus presentation and generation of many alternate forms. While these practical advantages afforded by computerisation may
provide a more uniformly administered test, uniform administration by itself does not necessarily bestow acceptable test reliability. Small statistical differences between groups of injured athletes and groups of control athletes are often cited as evidence of the sensitivity of neuropsychological tests. However, medical management decisions about concussed athletes are always made on a case-by-case basis. Evidence of differences between groups provides no information about the sensitivity of a test to change within individual athletes over time.

The ability to detect subtle changes in an individual’s neuropsychological test performance, like those commonly observed after concussion, is largely an issue of test reliability. Essentially, a reliable test is a test that contains very small amounts of measurement error. When using a reliable test repeatedly the clinician can be sure that any change in the measurement (test performance) reflects true change and not random variability. While all measuring devices contain error, tests that measure abstract constructs like cognition are prone to increased amounts of error. The greater the error in a test, the less sensitive it will be to detecting subtle change in individuals. Therefore, test sensitivity and test reliability and closely related.

When assessing the reliability and sensitivity of computerised neuropsychological tests for clinical and research use, potential users should inspect the psychometric properties of the test. Some important psychometric considerations that directly affect reliability, and therefore test sensitivity, are described below. If not designed with these issues in mind, computerised tests may be no more reliable or sensitive than the paper-and-pencil tests they are rapidly replacing.

**Number of observations**
Many neuropsychological tests make few observations of the individual’s behaviour when measuring their cognitive performance. For example, an athlete required to perform a computerised version of the Rey Auditory Verbal Learning Test (RAVLT) of memory may be distracted when responding to one of the fifteen trials of this test. This distraction was outside the control of athlete (random) and in no way reflects the athlete’s ‘normal’ level of performance, but the response time was abnormally slow as a consequence. If this was the only trial administered, the clinician might incorrectly infer that the athlete’s memory was abnormally slow. If multiple trials were administered, the effect of this erroneous score on the estimated average level of performance would diminish. This effect diminishes further as the number of trials on which the average is based increases. If the test required the individual to make only 5 responses, the mean is likely to be more affected by the single erroneous response than if the test required 50 responses. Thus, measurement error is reduced as the number of observations increases, and the reliability of the test increases because the effect of any error is diminished. Consequently the test is more likely to detect true changes in cognitive performance if they exist and this will facilitate more accurate clinical decision-making. To illustrate this important point a case example is described below.

Case example

Athlete X is a 22 year old elite Australian footballer. His past history includes one prior concussion approximately two years ago and no other significant history of head trauma or psychiatric illness. He was concussed following a collision with an opponent during the course of regular play and was immediately removed from the field and took no further part in the game. His initial symptoms included confusion, headache, dizziness and blurred vision. On review at day 1 post concussion he
described ongoing headache which was aggravated by activity and a general feeling of fatigue. All of his symptoms had resolved by day 4 post concussion.

At day 1 post-concussion, the performance of Athlete X on the CogSport psychomotor task was 1.72 standard deviations below baseline when all 75 baseline and 75 post-concussion responses were included in a z-score calculation (Table 1). This is a large decline in performance according to conventional statistical criteria \(^{10}\), and correlates well with Athlete X’s clinical presentation at day 1. No impairments relative to baseline were observed at day 4 post-concussion. Again, this correlates well with the clinical observations of symptom resolution at day 4, and indicates that the athlete’s brain function had returned to normal. On the basis of these cognitive and clinical findings, the treating physician allowed Athlete X to begin a graduated reintroduction to training at day 5 and resuming play the same week.

In Table 1, the results of a re-analysis of Athlete X’s cognitive data is presented. We randomly selected 5, 10, 20, 30 and 50 of the 75 total responses and calculated Athlete X’s mean and standard deviation of performance at both baseline and post-concussion assessments. Random response selection was accomplished using the Statistical Package for the Social Sciences (SPSS) version 10. Z-scores for each of these conditions were then calculated. When only five observations were included in the calculation, day 1 post-concussion performance was estimated to be 2.59 standard deviations slower than at baseline, a substantial decline in performance. This appears to have been caused by an increase in the day 1 post-concussion mean, probably influenced by outlying score/s. When ten observations were included in the calculation, change from baseline was estimated to be 0.90 standard deviations, a much more moderate performance decline. As more observations are included in the calculation (>30), the estimate begins to be reported more consistently as between 1.38 and 1.72 standard deviations. This is because a more
reliable estimate of actual performance is being gained as the number of
observations increases.

This case example demonstrates that estimates of change in neuropsychological test
performance between two testing occasions can vary greatly when relatively few
observations are made, even when those observations are made within the same
individual. The consequences for clinical decision making are substantial. For
example, had Athlete X been required to make only five responses at baseline and at
day 1 post-concussion, the clinician may have received results suggesting that the
athlete’s cognition was severely impaired after the concussion. In contrast, had ten
responses been required, the results would have suggested that the impairment was
very mild and perhaps even non-significant, as changes of at least 1 standard
deviation are generally required to infer clinically significant change in
neuropsychological test score\textsuperscript{10}. As a greater number of observations are included
in the analysis, estimates of change become stable and reliable clinical decisions can
be made.

Data produced by test

Another psychometric component of neuropsychological tests that may affect clinical
decision-making is the type of data produced by the test. Neuropsychological tests
that provide continuous data ranges (e.g., reaction time) are often very reliable and
sensitive to subtle changes in cognition, whereas tests that provide interval level data
(e.g., accuracy or number of correct/erroneous responses) often have poor
reliability\textsuperscript{9}. This is because the scale on which performance is measured directly
affects the ability to detect mild changes in test score.
Most computerized tests of reaction time have millisecond accurate timing, allowing one thousand possible levels of performance within every second of recording \(^6,^{11}\). It is therefore possible for a very mild change in average reaction time to be detected, as in the example of Athlete X above, where an average slowing of 130 ms from baseline was sufficient for significant results to be obtained. In contrast, many more complex neuropsychological tests are accuracy-based and require very few responses, thus limiting the number of possible levels of performance. For example, had Athlete X been assessed with the paper-and-pencil version of the RAVLT, there would have been very few possible levels of performance. This is because most healthy young adults perform in a restricted range between 10 and 15 on this test, and accurate reaction times cannot be recorded with paper-and-pencil tests. Further, it is not possible to make statistical decisions regarding the significance of change on tests where only a single score is obtained in the absence of an estimate of performance variability (i.e., standard deviation or standard error).

Analysis of recent studies utilising computerised neuropsychological tests in sports concussion and head injury illustrate this point. For example, Warden and colleagues\(^{12}\) observed significant slowing of simple reaction time (SRT) in a group of twelve US army cadet athletes tested 4 days post-concussion. Post-concussion performance of the same athletes on computerised measures of mathematical processing, working memory, matching to sample and digit symbol substitution were not significantly different from baseline. Close inspection of these tests reveals that the SRT test has psychometric properties that lend themselves to detection of mild impairment (i.e., many observations & continuous data range) whereas the other tasks do not possess such properties. For example, the output for the digit symbol substitution task is number of correct responses and as most individuals performed perfectly, this task appears to suffer from ceiling effects. Such ceiling effects limit the
ability of the test to detect mild cognitive changes, as even mildly impaired individuals will continue to perform well.

Similar findings to those of Warden and colleagues (2001) have been observed in previous studies of concussion and head injury. We propose that impairments were observed on the simple reaction time-based tests in these studies because such tasks generally have very good psychometric properties, not because sports concussion causes impairment in simple reaction time. Further, we propose that impairments were not observed on more complex computerised tests because they have relatively poor psychometric properties, not because the cognitive domains they are testing are unaffected by concussion. With relatively few alterations, it is possible to develop computerised neuropsychological tests of cognitive domains other than simple reaction with good psychometric properties. For example, the CogSport learning test requires that athlete to make fifty responses and provides both continuous (RT) and interval level (accuracy) data.

**Related issues**

A number of other factors may affect the psychometric properties of a neuropsychological test and therefore the test’s ability to detect the consequences of concussion and aid medical management of the athlete. A brief discussion of those issues specific to computerised tests follows.

It seems that every computerised neuropsychological test requires the individual to respond in a different way. For example, tactile responses can be made via the keyboard, touch screen, mouse or external response box while verbal responses can be made via a microphone or direct communication with the test administrator. Two separate issues arise here. The first is that accuracy of response timing is likely to
decrease as the amount of hardware between the response and the recording of the response increases. This may result in increased variability of recorded responses, independent of the variability within the individual. As stated above, increased variability will decrease the ability to detect mild changes in cognition. The second issue is that changing response modes between tests alters the complexity of the test. Many computerised neuropsychological test batteries require different responses for each test (e.g., keyboard response for test 1 and mouse response for test 2). This means that the individual must learn both the cognitive and response requirements of each test in order to perform at their best. With each change in the response requirements of a test the potential for erroneous responses, unrelated to the individual’s true cognitive state, increases. In order to ensure that only the cognitive components of the test are measured, it is important to ensure that the response requirements for each test are as uniform as possible.

In sports concussion, neuropsychological tests are administered serially. It is therefore important to ensure that tests are of equivalent difficulty, and assess the same cognitive function, each time they are administered. One of the often-stated advantages of computerised neuropsychological tests is that they allow the generation of many, or indeed almost infinite, alternate forms. While this is true, availability of alternate forms does not guarantee enhanced test reliability. If the alternate forms employed are not of equivalent difficulty, both reliability and the ability to identify changes in cognition may be compromised. The ability to generate equivalent alternate forms is affected by the type of stimuli employed. For example, it is difficult to ensure the equivalence of two language-based tasks using two distinct sets of words. This is because even common words have different rates of usage and because individuals themselves have different experience with language, related to their age, education level and history of language use (for example, many athletes
tested on language tasks where English words are used do not have English as their first language).

Conclusions

Although computerised neuropsychological tests have many potential advantages over paper-and-pencil tests, these advantages are not realised simply by the process of computerisation. Rather, computerisation introduces its own unique challenges for test designers. These include ensuring that an adequate number of responses are collected, ensuring that data collected has psychometric properties sufficient to allow detection of mild cognitive changes, ensuring that responses are collected in a uniform manner, and ensuring that alternate forms are of equivalent difficulty and assess equivalent cognitive domains. If not designed correctly, computerised neuropsychological tests may be no more reliable or sensitive than paper-and-pencil tests. Use of inappropriately designed computerised tests may compromise accurate clinical decision making and therefore jeopardise the health and safety of concussed athletes.
References


Acknowledgements

The authors would like to thank Dr Michael Makdissi and Dr Paul McCrory for their assistance with data collection for the case study presented.
**Table 1.** Mean, standard deviation and z-scores calculated for Athlete X with increasing numbers of observations. Mean and standard deviation data represents reaction times recorded in milliseconds.

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>Mean +/- standard deviation</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Day 1</td>
</tr>
<tr>
<td>5</td>
<td>247.6(73.5)</td>
<td>437.6(259.0)</td>
</tr>
<tr>
<td>10</td>
<td>263.4(111.6)</td>
<td>362.6(56.1)</td>
</tr>
<tr>
<td>20</td>
<td>240.5(92.4)</td>
<td>350.2(81.8)</td>
</tr>
<tr>
<td>30</td>
<td>231.5(78.0)</td>
<td>363.9(120.7)</td>
</tr>
<tr>
<td>50</td>
<td>237.7(86.6)</td>
<td>356.9(75.6)</td>
</tr>
<tr>
<td>75</td>
<td>235.6(73.6)</td>
<td>361.9(99.4)</td>
</tr>
</tbody>
</table>