OCCASIONAL PIECE

Psychometric issues associated with computerised neuropsychological assessment of concussed athletes

A Collie, P Maruff, M McStephen, D G Darby

Br J Sports Med 2003;**37**:556–559

Psychometric issues associated with computerised neuropsychological assessment in sports concussion are put forward. Issues critical to ensuring test reliability and sensitivity are discussed, with particular reference to how inappropriate test design can affect clinical decision making.

See end of article for authors' affiliations

Correspondence to:
Dr Collie, CogState Limited, 51 Leicester Street, Carlton, Vic 3053, Australia;
acollie@cogstate.com

Accepted 21 January 2003

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 medical management of individual athletes after concussion. 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.[2–5] The psychometric and practical limitations associated with these tests[6] has led to the development of a number of computerised neuropsychological test batteries.[7 8] This brief article introduces some of the psychometric problems 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, and allows randomised stimulus presentation and generation of many alternative forms. Although 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.[2–5] 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 a subject's neuropsychological test performance, such as those commonly observed after concussion, is largely an issue of test reliability.[6] Essentially, a reliable test is one that contains very small amounts of measurement error.[9] 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. Whereas all measuring devices contain error, tests that measure abstract constructs such as cognition are prone to more error. The greater the error in a test, the less sensitive it will be to subtle change in individual subjects. Therefore test sensitivity and test reliability are 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 in mind, computerised tests may be no more reliable or sensitive than the paper and pencil tests that they are rapidly replacing.

NUMBER OF OBSERVATIONS

Many neuropsychological tests make few observations on the subject'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 15 trials of this test. This distraction is outside the control of the athlete (random) and in no way reflects the athlete's ''normal'' level of performance, but the response time is abnormally slow as a consequence. If this is the only trial administered, the clinician may incorrectly infer that the athlete's memory is 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 subject to make only five 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 about two years ago and no other notable 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 on day 1 after concussion, he described a continuing headache, which was aggravated by activity, and a general feeling of fatigue. All of his symptoms had resolved by four days after the concussion.

On day 1 after the concussion, the performance of the athlete 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 the athlete's clinical presentation on day 1. No impairments relative to baseline were observed four days after the concussion. Again, this correlates well with the clinical observations of symptom resolution on 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 on day 5 and to resume play the same week.

In table 1, the results of a reanalysis of athlete X's cognitive data are presented. We randomly selected 5, 10, 20, 30, and 50 of the 75 total responses and calculated the mean (SD) of performance at the baseline assessment and after concussion. 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, performance on day 1 after concussion 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 mean, probably influenced by outlying score/s. When 10 observations were included in the calculation, the 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 shows 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 on day 1 after concussion, the clinician may have received results suggesting that the athlete's cognition was severely impaired after the concussion. In contrast, had 10 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.[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—for example, reaction time—are often very reliable and sensitive to subtle changes in cognition, whereas tests that provide interval level data—for example, accuracy or number of correct/erroneous responses—often have poor reliability.[9] This is because the scale on which performance is measured directly affects the ability to detect mild changes in test score.

Most computerised tests of reaction time have millisecond accurate timing, allowing 1000 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 milliseconds 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, 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 on the significance of change on tests in which only a single score is obtained in the absence of an estimate of performance variability—that is, standard deviation or standard error.

Analysis of recent studies using computerised neuropsychological tests in sports concussion and head injury

Table 1 Mean (SD) and z scores calculated for athlete X with increasing numbers of observations

	Mean (SD)			z Score	
No of observations	Baseline	Day 1	Day 4	Day 1	Day 4
5	247.6 (73.5)	437.6 (259.0)	190.8 (66.6)	2.59	–0.78
10	263.4 (111.6)	362.6 (56.1)	225.4 (17.6)	0.90	–0.34
20	240.5 (92.4)	350.2 (81.8)	233.5 (102.7)	1.19	–0.08
30	231.5 (78.0)	363.9 (120.7)	233.7 (64.1)	1.66	0.03
50	237.7 (86.6)	356.9 (75.6)	237.1 (79.4)	1.38	–0.01
75	235.6 (73.6)	361.9 (99.4)	232.8 (66.2)	1.72	–0.04

Mean (SD) data represent reaction times recorded in milliseconds.

illustrate this point. For example, Warden and colleagues[12] observed significant slowing of simple reaction time in a group of 12 US army cadet athletes tested four days after 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 simple reaction time test has psychometric properties that lend themselves to detection of mild impairment—that is, many observations and 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 subjects 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 athletes will continue to perform well.

Similar findings to those of Warden and colleagues[12] have been observed in previous studies of concussion and head injury.[7 11 13–17] 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 the athlete to make 50 responses and provides both continuous (reaction time) and interval level (accuracy) data.[18]

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 these matters specific to computerised tests follows.

It seems that every computerised neuropsychological test requires the subject to respond in a different way. For example, tactile responses can be made using the keyboard, touch screen, mouse, or external response box, and verbal responses can be made using a microphone or direct communication with the test administrator. Two separate problems 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 subject. As stated above, increased variability will decrease the ability to detect mild changes in cognition. The second problem is that changing response modes between tests alters the complexity of the test. Many computerised neuropsychological test batteries require different responses for each test—for example, keyboard response for test 1 and mouse response for test 2. This means that the subject 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 subject's true cognitive state, increases. 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, alternative forms. Although this is true, availability of alternative forms does not guarantee enhanced test reliability. If the alternative forms used are not of equivalent difficulty, both reliability and the ability to identify changes in cognition may be compromised. The ability to generate equivalent alternative forms is affected by the type of stimuli used. 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 in which 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, that the data collected have psychometric properties sufficient to allow detection of mild cognitive changes, that responses are collected in a uniform manner, and that alternative 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.

ACKNOWLEDGEMENTS

We thank Dr Michael Makdissi and Dr Paul McCrory for their assistance with data collection for the case study presented.

Authors' affiliations

A Collie, P Maruff, M McStephen, D G Darby, CogState Limited, Carlton, Australia
A Collie, M McStephen, D G Darby, Centre for Neuroscience, University of Melbourne, Australia
P Maruff, Department of Psychology, La Trobe University, Melbourne, Australia

REFERENCES

1 **Aubry M**, Cantu R, Dvorak J, *et al.* Summary and agreement statement of the first International Conference on Concussion in Sport, Vienna 2001. *Br J Sports Med* 2002;**36**:6–10.
2 **Butler R**. Neuropsychological investigation of amateur boxers. *Br J Sport Med* 1994;**28**:187–90.
3 **Collins MW**, Grindel SH, Lovell MR, *et al.* Relationship between concussion and neuropsychological performance in college football players. *JAMA* 1999;**282**:964–70.
4 **Maddocks D**, Saling M. Neuropsychological deficits following concussion. *Brain Injury* 1996;**10**:99–103.
5 **Matser JT**, Kessels AG, Jordan BD, *et al.* Chronic traumatic brain injury in professional soccer players. *Neurology* 1998;**51**:791–6.
6 **Collie A**, Darby DG, Maruff P. Computerised cognitive assessment of athletes with sports related head injury. *Br J Sports Med* 2001;**35**:297–302.
7 **Makdissi M**, Collie A, Maruff P, *et al.* Computerized cognitive assessment of concussed Australian rules footballers. *Br J Sports Med* 2001;**35**:354–60.
8 **Erlanger D**, Saliba E, Barth JT, *et al.* Monitoring resolution of post-concussion symptoms in athletes: preliminary results of a Web-based neuropsychological test protocol. *Journal of Athletic Training* 2001;**36**:280–7.
9 **McCaffrey RJ**, Duff K, Westervelt HJ. *Practitioner's guide to evaluating change with neuropsychological assessment instruments.* New York: Kluwer Academic/Plenum Publishers, 2000.
10 **Zakzanis KK**. Statistics to tell the truth, the whole truth and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Arch Clin Neuropsychol* 2001;**16**:653–67.

11 **Bleiberg J**, Garmoe WS, Halpern EL, *et al*. Consistency of within-day and across-day performance after mild brain injury. *Neuropsychiatry, Neuropsychol Behav Neurol* 1997;**10**:247–53.

12 **Warden DL**, Bleiberg J, Cameron KL, *et al*. Persistent prolongation of simple reaction time in concussion. *Neurology* 2001;**57**:524–6.

13 **Stuss DT**, Stethem LL, Hugenholtz H, *et al*. Reaction time after head injury: fatigue, divided and focused attention, and consistency of performance. *J Neurol, Neurosurg Psychiatry* 1989;**52**:742–8.

14 **Stuss DT**, Pogue J, Buckle L, *et al*. Characterization of stability of performance in patients with traumatic brain injury: variability and consistency on reaction time tests. *Neuropsychology* 1994;**8**:316–24.

15 **Hugenholtz H**, Stuss DT, Stethem LL, *et al*. How long does it take to recover from a mild concussion? *Neurosurgery* 1988;**22**:853–8.

16 **Van Zomeren AH**, Deelman BG. Long-term recovery of visual reaction time after closed head injury. *J Neurol Neurosurg Psychiatry* 1978;**41**:452–7.

17 **Van Zomeren AH**, Deelman BG. Differential effects of simple and choice reaction after closed head injury. *Clin Neurol Neurosurg* 1976;**79**:81–90, 79;81–90).

18 **Collie A**, Maruff P, Makdissi M, *et al*. CogSport: reliability and correlation with conventional cognitive tests used in post-concussion medical examinations. *Clin J Sport Med* 2003;**13**:28–32.