This is a repository copy of *Expansion of cognitive testing for off-field concussion screening in elite rugby players: a cohort study*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/175305/

Version: Accepted Version

Article:
Fuller, G.W. orcid.org/0000-0001-8532-3500, Gardner, A., Tucker, R. et al. (7 more authors) (2021) Expansion of cognitive testing for off-field concussion screening in elite rugby players: a cohort study. Journal of Science and Medicine in Sport, 24 (12). pp. 1204-1210. ISSN 1440-2440

https://doi.org/10.1016/j.jsams.2021.06.002

Article available under the terms of the CC-BY-NC-ND licence (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Abstract

Objectives

Current off-field concussion screening instruments have sub-optimal accuracy and additional testing domains may be necessary to detect the full spectrum of concussion presentations. This study aimed to determine if additional cognitive tests add utility to off-field screening for sport-related concussion.

Design

Reproducibility and diagnostic accuracy cohort studies were performed in the 2017 and 2018 seasons of the Super Rugby competition, conducted in Argentina, Australia, Japan, New Zealand, and South Africa.

Methods

Abridged versions of Stroop (score, time), Spatial Memory (score, failed trials), and Trail Making Trial-B (time, errors) cognitive tests, modified for off-field use, were examined. Players performed baseline testing prior to each season. Cases undergoing off-field screening as part of the World Rugby Head Injury Assessment Process underwent evaluation with the same cognitive tests during competition matches. Agreement between repeated pre-season baseline measurements, and the diagnostic accuracy of off-field testing against a clinical reference standard of concussion, were evaluated.

Results

Data were available for repeated preseason baseline testing in 644 players, and 100 cases undergoing off-field concussion screening. There was little individual agreement across pre-season baseline assessments for all tests (Lin's correlation and Gwets AC1 coefficients ranging between 0.2 and 0.3). There was significantly worse performance for the time taken to complete the modified Stroop Test in concussed players undergoing off-field screening, compared to non-concussed players (median time 21.1 v 18.4 seconds, $p<0.01$; area under the receiver operating characteristic curve 0.7 (AUROC, 95%CI

0.5-0.8). Other cognitive measures did not discriminate between injured and un-injured players, with no-statistically significant differences in distribution medians (p=0.6-0.9) and AUROC values close to 0.5.

Conclusions

The time taken to perform a modified Stroop Test may provide additional diagnostic accuracy if added to current off-field concussion screening tools. Abridged Spatial Memory and Trail Making Trial-B tests did not discriminate between concussed and non-concussed players.

Key words

Concussion, screening, sensitivity, specificity, agreement, Rugby

Practical implications:

1. Current off-field concussion screening tools have sub-optimal accuracy. Stroop, Spatial memory, and Trail Making Trial-B tests are commonly used cognitive tests. Abridged versions, suitable for side-line use, could improve off-field detection of concussion if added to existing screening tools.

2. The time taken to perform a modified Stroop Test may provide additional diagnostic accuracy if added to current off-field concussion screening tools.

3. Modified Spatial Memory and Trail Making Trial-B tests did not discriminate between injured and un-injured players.

4. Reproducibility of repeated testing was low for all studied cognitive tests, suggesting limited benefit from baseline testing.

Introduction

Concussion is a common and high profile injury in contact and collision sports.[1] Continuing to play when concussed could worsen symptoms, delay recovery, and increase the risk of further injuries.[2] Elite sports, following recommendations from the Berlin Concussion Consensus Statement, have therefore introduced systems to identify and manage head impact events with the potential for concussion during matches.[3] These typically involve off-field (often termed side-line) testing to detect suspected concussion following suspect head impact events.[3]

In elite Rugby, the World Rugby Head Injury Assessment off-field screening tool (HIA-1) is used to assess players following identification of a meaningful impact event, without overt signs of concussion.[4, 5] This instrument includes sub-tests assessing cognition (orientation, memory, and concentration), balance, and symptomatology. It takes 10 minutes to administer and has demonstrated a sensitivity of 77% and specificity of 87% for the diagnosis of concussion. The relatively low reported sensitivity of the HIA-1 off-field tool may offer the opportunity to reduce the number of false negative cases through more detailed testing.

Neuropsychological tests, including computerized neurocognitive tests, are used for diagnosis, assessment of severity, and monitoring of recovery in neurological diseases e.g. traumatic brain injury, hepatic encephalopathy.[6, 7] The Trail Making, Stroop, and Corsi block tapping tests are established neuropsychological tests that have been studied in moderate and severe traumatic brain injury,[8-10] Although not well studied in acute mild head injury, they evaluate a range of cognitive functions, and involve different brain pathways that could conceivably be affected by concussion.[6, 7]

These detailed tests are not suitable for off-field application during sporting competitions due to time constraints. However, it might be hypothesised that shortened versions, retaining the underlying

construct, could be effective as additional novel screening tests for concussion; although there is currently an absence of evidence for their diagnostic accuracy following head impact events in sports.[11]

This study therefore aimed to determine if abridged versions of Trail Making, Stroop and Corsi block tapping tests, modified for off-field use, could be useful additions to screening tools for sport-related concussion. Specific objectives were to: describe the distribution, reproducibility and normative values of baseline cognitive test values; describe the distributions of cognitive test scores in concussed and non-concussed players; and evaluate diagnostic accuracy in Rugby Union players undergoing off-field concussion screening.

Methods

Prospective reproducibility and diagnostic cohort studies were conducted in the male, adult, elite-level Rugby Union 'Super Rugby' competition players over 2 seasons between 2017 and 2019. The source population consisted of players rostered on Australian (5 teams), New Zealand (5), South African (6), Argentinian (1) and Japanese (1) elite level rugby union teams (n=644). The study populations comprised rostered players undergoing pre-season baseline testing; and those undergoing an off-field HIA-1 concussion screening tool assessment after identification of a meaningful head impact event during competitive matches with the potential to cause concussion, but where the consequences were unclear i.e. the possibility of concussion in the absence of overt manifestations, e.g. dangerous mechanism. The indications for off-field screening assessments (termed 'criteria 2'), and details of the HIA-1 screening tool, are provided in the web appendix (Table A1 and A2).[4]

According to the World Rugby HIA process, players demonstrating clear signs of concussion (e.g. loss of consciousness, tonic posturing, ataxia, seizures) during matches are immediately and permanently removed from the remainder of the match, without undergoing further off-field concussion screening. These players were therefore outside of the scope of current study. The indications for immediate and permanent removal (termed 'criteria 1'), and further details of the HIA process, are provided in the web appendix.[4]

The Trail Making, Stroop, and Corsi block tapping cognitive tests were chosen for evaluation based on the following factors: recognition as established and validated neurocognitive tests;[6, 7] previous use in traumatic brain injury;[8-10] availability, feasibility and practicality of application as a brief screening test; and complementation of cognitive domains assessed within the current HIA-1 off-field screening tool. In elite Rugby, the HIA-1 concussion screening tool is administered during a 10-15 minute off-field assessment period. As each of the chosen cognitive tests take approximately 5 minutes each to

complete, it was not possible to study the traditional versions alongside the existing HIA-1 screening tool in the time available. Previously unvalidated, abridged (taking approximately 1-2 minutes each to administer) versions of the Trail Making, Stroop, and Corsi block tapping tests were therefore developed by CSx Systems (CSx Systems, Auckland, New Zealand) to allow off-field screening evaluation as described below:

- **Trail Making Test:** The Trail Making Test (TMT) provides information on visual search, scanning, speed of processing, mental flexibility, and executive functions.[12, 13] The TMT consists of two parts. TMT-A requires an individual to draw lines sequentially connecting 25 encircled numbers distributed on a sheet of paper. Task requirements are similar for TMT-B except the person must alternate between numbers and letters (e.g., 1, A, 2, B, 3, C, etc.).[13] This study only used the TMT-B test, which was presented on a tablet and required the participant to complete the test by tapping on the target stimulus in the order outlined above. Total time to completion and errors were recorded.

- **Stroop test:** The Stroop colour-word interference task measures selective attention, mental flexibility, and inhibitory control.[14, 15] Conventionally subjects are required to read three different tables as fast as possible. Two of them represent the 'congruous condition' in which participants are required to read names of colours and name different colour patches. In the third 'incongruent' table, colour-words are printed in an inconsistent colour (e.g. the word "red" is printed in the colour green), and participants are required to name the colour instead of reading the word. Unlike the traditional version of the Stroop, in this study the test taker was not required to perform the priming subtests (word reading or colour naming) prior to completing the interference subtest (inconsistent colour-words). Study subjects were presented with 10 colour-words of varying colour congruency on a tablet and required to respond by

identifying the correct colour displayed. The time taken to perform the task and score were recorded.

- **Corsi block tapping test:** The corsi block tapping test assesses visuo-spatial working memory and spatial attention.[13] The traditional forward version consists of nine cubical blocks irregularly positioned on a board. An examiner taps the blocks starting with sequences of two cubes. The subject has to reproduce a given sequence by tapping the blocks in the same sequence. If a certain proportion of the sequences is reproduced correctly (usually 1/2, 2/3, or 3/5 of the trials per sequence length), the sequence length increases by one item. The procedure ends when the number of wrong reproductions exceeds the proportion of admissible errors per length.[16, 17] In this study a brief spatial memory test, modifying the forward Corsi block test, was investigated. A symmetrical 3 x 3 grid of 9 objects (schematic flowers) was presented on a tablet. One item was randomly and sequentially highlighted at a time and respondents were asked to repeat the order that these items appeared by tapping the corresponding objects on the grid. Five trials were performed with the span (that is, the length of the pattern sequence) increasing after successful completion of a sequence, and decreasing after failures, in the range from 3 (minimum span) to 7 (maximum span). The number of failed trials and score (calculated from number of successful trials and spans, range 5 – 400) of the modified Corsi block test was recorded.

Pictures of each cognitive test are shown in the web appendix (Figure A1). Players in participating Super Rugby teams underwent baseline cognitive test evaluation after completion of a routine training session. Post-exertional baseline testing was chosen to reflect the conditions present in off-field concussion screening conducted during competitive matches. Testing was conducted in a standardised format by the team doctor. This predominantly occurred prior to each season but was also repeated in

the event of transfer to another team. Thus, depending on team rosters, players could undergo baseline testing: (i) once if participating in the first season of the study period only; (ii) twice if playing in the Super Rugby competition during both study seasons; or (iii) more than twice if transferring to different teams in the competition.

Participants requiring off-field HIA-1 concussion screening during competitive matches then underwent assessment with the cognitive index tests during a 15 minute off-field medical room assessment administered by the team doctor. Cognitive index test performance did not influence return to play decisions, with tests implemented non-operationally after completion of the HIA-1 screening tool. Team doctors could observe gross cognitive test performance, but were blinded to the final results (i.e. total errors, score, time).

Following an off-field concussion screen, repeat clinical assessments were made by the team doctor post-match (HIA-2 assessment) and after 48 hours rest (HIA-3 assessment), to determine a final clinical diagnosis of concussion.[18] The HIA-2 assessment consists of a clinical evaluation including the Sports Concussion Assessment Tool 5th Edition (SCAT-5)instrument.[19] The HIA-3 assessment comprises a clinical evaluation, supported by an expanded SCAT-5 symptom checklist, a cognitive assessment (typically a computerised neuro-cognitive tool such as CogSport) and a balance assessment using the balance error scoring system and tandem gait balance tests.[19, 20]

HIA process data are routinely recorded at the point of assessment by doctors using a tablet based, web-hosted platform, with standardised data collection forms (CSx Systems, Auckland, New Zealand). All cognitive tests were administered using software applications embedded within the CSx data collection system. Spanish and Japanese language-specific versions of the Stroop and TMT-B tests were available for the Argentinean and Japanese teams.[21-23] Data was subsequently uploaded to the World Rugby HIA database. HIA assessment forms, from each of the 3 HIA process stages, are linked deterministically using unique player identifiers. Competition coordinators are responsible for data quality and collection

of outstanding information. Any missing CSx outcome data was collected by direct communication with team doctors.

Analyses proceeded in 3 stages. Firstly, player flow through the study was described. Secondly, pre-season baseline results were examined. The distributions of pre-season baseline cognitive test scores were examined using descriptive statistics. Only one baseline value was included per player, with the first result used if repeated results were available. Normative ranges were then determined for each cognitive test. Cut-offs were selected based on distribution percentiles consistent with previous SCAT normative value studies and followed conventions used in cognitive assessments (e.g. Wechsler intelligence quotient classifications).[24-26] The below average cut-off was defined as close as possible to the $25^{th}/75^{th}$ percentile ranks. Unusually low scores corresponded to the $10^{th}/90^{th}$ percentile ranks, and extremely low scores aimed for the $2^{nd}/98^{th}$ percentile ranks. The reproducibility of cognitive tests was then evaluated in players with repeated pre-season baseline results. In the event of multiple baseline evaluations, the first two repeated measurements were examined. Scatter plots and frequency histograms were used to visually compare the relationship between repeated measurements. Bland-Altman Limits of Agreement and Lin's concordance coefficient (continuous time and score variables); and two-way bar charts, Cohen kappa using linear weights, and Gwets AC1 statistic (discrete score, test errors, and failed trials variables), were then used to quantify agreement.

Thirdly, cognitive test performance in players undergoing off-field concussion screening during games was investigated. The distribution of test results was compared between concussed and non-concussed players graphically and with Mann Whitney U tests. The sensitivity, specificity, and area under the receiving operating characteristic curve (AUROC) were calculated for the cognitive index tests and the HIA-1 screening tool against the reference standard of a final clinical diagnosis of concussion. The primary threshold for cognitive index test positivity was the empirically derived normative threshold for poor performance ($10^{th/}90^{th}$ percentile ranks of baseline test result distribution, described above).

Secondary exploratory cut-points were explored according to any deterioration from the most recent baseline value and empirical thresholds derived using Youden's J statistic.

Buderer's formula was used to estimate sample size with the following assumptions from previous HIA data: a prevalence of concussion of 40% in players with significant head impact events requiring HIA-1 concussion screening assessment; a target sensitivity of 90%; and a target specificity of 80% for cognitive index tests to identify concussion.[4, 5, 27] To achieve a desired precision of 10% for the sensitivity estimate would require a sample size of 88 players undergoing off-field concussion screening assessments.

Available case analyses were performed with the sample size determined by the number of players with complete data for each analysis. Statistical analyses were carried out in Stata version 13.1 (StataCorp, College Station, USA) with a conventional significance level (α) of 0.05 used. Ethical approval was provided by the University of Sheffield (Reference Number 017814). All players provided informed consent for participation prior to the start of the season. CSx provided an anonymised data set for analysis with player identifiers removed.

Results

Of the 644 players rostered on Super Rugby teams during the study, all were male with a median age of 24 years (IQR 22-28, range 18-37) and a median professional Rugby experience of 4 years (IQR 3-7, range 1-19). The majority of participants (85%) spoke English as their first language. Most players had been diagnosed with at least one previous concussion (64%, median 1, IQR 0-2, range 0 – 10). Players had predominantly completed education to secondary school level (82%, with a median of 13 years education (IQR 12-14, range 7-20).

Baseline testing was performed on 896 occasions in the 15 participating teams over the 2 seasons of the study. This testing was performed in 643 of 644 rostered players (Figure 1, once in 409 players, repeated twice in 215 players, repeated thrice in 19 players). Data were missing due to incorrect or absent baseline testing in a small number of cases (1.6% of total number of possible baseline tests across all index tests, testing occasions, and players, n=5,286/5,376).

The cognitive tests generally exhibited a unimodal, positively skewed distribution, with a long tail; indicating the majority of players performing well with a minority demonstrating increasingly poor performance (web appendix Table A3, web appendix Figure A2). The normative classification ranges for each cognitive test are presented in in the web appendix (web appendix Table A4).

Distributions for repeated baseline measurements were similar for all cognitive tests, with inter-athlete baseline results before season 2 of the study mirroring the overall pattern of those from season 1 (web appendix Figures A3). In contrast, there was little intra-athlete agreement for each participant with scatter plots showing a lack of correlation between repeated measurements (web appendix Figure A4). Lin's correlation coefficient (ρc) for TMT-B time, Stroop test time, and spatial memory test score ranged between 0.21 and 0.33 (Table 1), with Bland Altman plots showing large 95% limits of agreement exceeding plausible clinically acceptable interchangeability thresholds (Table 1, web appendix Figure

A5). Agreement between repeated baseline results for TMT-B errors, Stroop test score, and spatial memory test failed trials was low, with exact agreement evident in only 37.0%, 22.8% and 49.1% of players (Table 1), respectively. The weighted kappa statistic also indicated poor agreement beyond that expected for each cognitive test (Table 1): TMT-B errors 0.01 (95%CI 0.0-0.09); Stroop test score 0.20 (95%CI 0.12-0.28); Spatial memory test failed trials 0.11 (95%CI 0.01-0.21). Gwets AC1 coefficients ranged between 0.2 and 0.3 (Table 1).

A total of 211 consecutive meaningful head impact events were detected in 165 individual players (recurrent events occurring in 33 players, ranging from 24 players with 2, to 2 players with 5 incidents) during competitive Rugby matches over the study period. Of these, 84 incidents (in 74 players) were associated with overt signs or symptoms of concussion requiring immediate and permanent removal from play. The remaining 127 incidents (in 106 players), where the consequences were unclear, underwent off-field concussion screening comprising the study population. 100 players were assessed with the cognitive index tests and were included in the final study sample (Figure 1). Of these cases, 33 players were diagnosed with concussion giving a prevalence of 33.0% (95% CI 24.4-42.9%).

There was significantly worse performance in the time taken to complete the Stroop Test in concussed, compared to non-concussed players (median time 21.1 v 18.4 seconds, $p<0.01$, AUROC 0.66 (95% CI 0.54-0.76), n=100, Table 2, web appendix Figure A6). Using the normative threshold (web appendix Table A4) as the cut-point for abnormal performance demonstrated low sensitivity, but high specificity, of 15.2% (95%CI 5.1-31.9%) and 91.0% (95%CI 81.5-96.6%) respectively (Table 2). Using any worsening from baseline as the threshold for test positivity offered the opposite pattern of higher sensitivity with lower specificity with values of 75.0% (95%CI 53.4-90.2%) and 50.0% (95%CI 34.6-65.4%) respectively (web appendix Table A5). An optimal cut-point of 19.8 seconds was indicated by the Youden method. Sensitivity at this threshold was 66.7% (95%CI 48.2-82.0%) with a specificity of 68.7% (95%CI 56.2-79.4%).

The existing HIA01 screening tool in this cohort had a sensitivity of 78.8% (95%CI 61.1-91.0%) and specificity of 64.2% (51.5-75.5%, n=100) for the detection of concussion in included players. Adding time to complete the Stroop test to the HIA01 screening tool using its normative threshold resulted in the detection of 2 additional concussion cases, at the cost of 4 false positives (combined sensitivity of 84.8%, specificity 58.2%, AUROC 0.72). Using empirical Youden or baseline thresholds in a combined HIA01 screening tool detected additional concussion cases (3 and 3 respectively) but at the cost of a high number of false positives (11 and 17 respectively).

Performance was similar regardless of injury status for all other cognitive index tests, with no-statistically significant differences in distributions (Mann-Whitney U tests, p=0.58-0.98, Table 2) or frequency histograms (web appendix Figure A6). AUROC values were close to 0.5 indicating minimal ability to differentiate between concussed and non-concussed players (web appendix Figure A7). Diagnostic accuracy was poor for these other cognitive index tests regardless of whether normative (sensitivity 12.1% - 24.2%; specificity 65.7% - 89.6%) or baseline (sensitivity 29.1% - 75.0%; specificity 50.0% - 73.8%) thresholds for positivity were used; with fewer additional concussions detected at the cost of more false positive cases (Table 2 and web appendix Table A5).

Discussion

There are several important findings from the present study. Firstly, abridged Stroop (score, time), Spatial memory (score, failed trials), and TMT-B (time, errors) tests demonstrated poor intra-athlete agreement, with players often having markedly different results across repeated pre-season baseline assessments. Secondly, in players undergoing off-field screening assessments, there was significantly worse performance for the time taken to complete the abridged Stroop test in concussed players compared to non-concussed players (median time 21.1 v 18.4 seconds, $p<0.01$; AUROC 0.66 (95% CI 0.54-0.76; n=100). Adding this test to the HIA01 screening tool, using a normative cut-point, detected a small number of additional concussion cases, at the cost of relatively few false positives. Finally, errors in the abridged Stroop test, and abridged Spatial memory and TMT-B cognitive index tests, did not appear to discriminate between injured and un-injured players, with no-statistically significant differences in distribution medians ($p=0.58$-0.98) and AUROC values close to 0.5.

The time taken to complete a 10-word Stroop colour-word interference task was the only studied cognitive measure that discriminated between concussed and non-concussed players undergoing off-field screening assessments. Processing of the Stroop task is performed by the anterior cingulate cortex and the dorsolateral prefrontal cortex.[11,12] Impairment in both of these cerebral regions has been observed in human MRI imaging studies, and animal models, following mild traumatic brain injury.[28] Although evidence is limited in concussion and mild head injury,[29] studies have shown worsened performance in the traditional Stroop task in moderate and severe traumatic brain injury.[8] An increased cognitive interference effect (delayed reaction time due to mismatch in stimuli) has also been found in other neurological disorders including dementia, attention-deficit hyperactivity disorder, and Parkinson's disease strengthening the biological plausibility of a detectable deficit in concussion.[30-32] Further modification of the Stroop test to include other sensory modalities and variables, for example testing the reverse Stroop effect (reading a colour-word printed in the congruent colour, an incongruent

colour, or a neutral colour), could conceivably magnify this effect. The observed lack of utility of the abridged Spatial Memory and TMT-B tests suggests that either the associated cognitive functions are not commonly impaired acutely post-concussion, or that the studied tasks were not sufficiently demanding.

Current concussion screening tools consist of multiple sub-tests conducted in parallel, in which all tests need to be normal to give a negative result and any abnormal finding results in a positive test. In this testing paradigm, as further sub-tests are added sensitivity will increase, and specificity fall, dependent on the performance and overlap of individual sub-tests. The sensitivity and specificity of these constituent sub-tests can also be varied in opposite directions as the cut-point for test positivity is altered; with the optimal trade-off dependent on how false negative and false positive cases are valued. The World Rugby HIA01 multi-modal concussion screening tool has previously shown moderate sensitivity and high specificity.[4, 5] The key aim when expanding off-field concussion screening is therefore to select sub-tests and test thresholds that detect additional concussion cases, without an unacceptable increase in false positives. The individual sub-tests currently included in the HIA01 screening tool have previously demonstrated AUROC values between 0.51 and 0.73.[4] The time to complete the abridged Stroop test AUROC of 0.66 compares favourably to these. Using normative thresholds for Stroop test time demonstrated high specificity, and therefore if sufficiently sensitive, offers promise to detect extra concussion cases at the cost of very few extra false positives. Using baseline thresholds, or the empirically derived cut-point using Youden's index, for the abridged Stroop test would detect a small number of additional concussions but would likely result in an unacceptable loss of specificity.

Baseline test results are commonly used in off-field screening assessments to aid interpretation of post-injury scores, although they are not considered mandatory in the Berlin Concussion Consensus Statement.[3] The studied cognitive tests were imprecise with significant natural variation apparent with

repeated testing. The relatively high intra-athlete variability noted for each cognitive test between each pre-season measurement, is similar to that reported previously for serial SCAT assessments,[33] and would seem to limit the utility of baseline testing. Worse results than baseline would be frequently expected secondary to random natural variation in test performance rather than underlying cognitive impairment. More importantly, the lack of precision of the cognitive tests (closeness of the repeated measurements) will undermine test performance, even if relatively accurate (closeness of the measurements to a specific value). The lack of agreement observed with repeated Stroop test time measurements, implies high measurement error for the task and an inability to attribute change to a concussion unless there change of large magnitude. This could therefore support the use of normative, rather than baseline thresholds.

The source population of professional Rugby players should ensure that these results are largely generalisable throughout elite Rugby Union competitions. However, previous research has suggested that team doctors often use clinical judgement to interpret off-field screening subtest results with improved overall discrimination for concussion compared to strict application of baseline or normative thresholds.[4] Cognitive tests were implemented non-operationally and it is therefore possible that modified Stroop test performance could differ if applied in actual practice. External validity to the elite level of other sports with different frameworks for evaluating head impact events is less certain, although cognitive index test administration should largely be independent of context-specific factors such as test setting (on field v pitch-side v medical room) or equipment (boots v skates). Results could also differ if alternative formats of the abridged tests are administered, for example using paper-based forms rather than a tablet application.

This study provides a pragmatic evaluation of novel cognitive tests for rapid sports related concussion screening on the off-field and has several strengths. The cognitive tests and reference standard were applied to all included participants with no potential for partial or differential verification biases. The

reference standard was determined after serial standardised examinations by experienced team doctors reducing the risk of differences in diagnostic thresholds and misclassification. Furthermore, blinding, and sequential outcome assessments prevented test and diagnostic review biases.

Conversely, there are a number of limitations which could challenge internal validity. Firstly, as a real-life study using routinely collected clinical data there were inevitably missing data with the potential for selection bias. However, this is likely of less magnitude than performing a two-gate 'case-control' study design common in this research field. Secondly, there is the possibility of incorporation bias. During normal sports medicine practice the team doctor predominantly conducts both off-field screens and subsequent diagnostic assessments. Although blinded to the final cognitive index test results, it is possible that team doctors incorporated a global impression of screening test performance into their overall diagnostic assessment. Thirdly, the diagnosis of concussion is inherently subjective and ill-defined, as acknowledged in the Berlin concussion consensus document. Imperfect diagnosis of concussion could therefore lead to misleading accuracy metrics. The extent of any systematic error would be dependent upon the frequency of misclassifications and the degree of correlation between cognitive test and reference standard errors. Fourthly, language-specific tests were not available for all participants and some players may have completed the cognitive tests in a non-native language. In mitigation, the vast majority of players were fluent in English and previous studies examining differences in the Stroop effect in bilingual individuals are inconclusive. Moreover, language-specific versions were available for use in Spanish and Japanese teams, with similar baseline results observed across different languages (web appendix Figure A8). Fifthly, although the cognitive index tests were implemented in a uniform fashion, we did not investigate inter-observer variability in performance between team doctors. Finally, atypical testing paradigms were used for each investigated cognitive index test. The observed lack of diagnostic utility could reflect the ineffectiveness of these non-validated testing approaches or sub-optimal administration by non-experts. However, the construct for the abridged tests remains

consistent with the traditional versions. Furthermore, the formal versions of these tests are not practical for brief off-field concussion screening and testing by non-neuropsychologists reflects the reality of real-world sports medicine. As stated in the Berlin concussion consensus statement 'abbreviated testing paradigms are designed for rapid sports related concussion screening on the sidelines and are not meant to replace a comprehensive neurological evaluation'.[3]

This is the first study to evaluate the Stroop, Spatial memory and TMT-B tests for off-field sports-related concussion screening. However, there is a larger literature examining normative values for the TMT-B and evaluating the performance of the cognitive tests in patients with confirmed traumatic brain injury. The distribution of TMT-B times observed in the current study are significantly quicker than previously published normative values (median 17 v 47 seconds). This is likely to reflect different test taking conditions using a tablet application compared to traditional administration with paper and pencil. In contrast to TMT-B time, a cut-point of 3 errors has been suggested in previous studies,[34] which appears consistent with the normative range derived from the current investigation. A number of studies have demonstrated impaired performance of the Stroop, Spatial memory, and TMT-B in traumatic brain injury patients,[35,29,30] however the relevance of these to the milder sub-set of concussed patients is uncertain.

Conclusions

The time to perform an abridged Stroop test, modified for off-field use, may provide additional diagnostic accuracy if added to current off-field concussion screening tools. Reproducibility of repeated testing was low, suggesting limited benefit from baseline testing, and supporting use of a normative threshold for test positivity. Modified Spatial memory and TMT-B tests did not discriminate between concussed and non-concussed players and do not appear to have utility in off-field screening assessments.

References

1. McCrea HJ, Perrine K, Niogi S, Hartl R. Concussion in sports. *Sports Health.* 2013; 5(2):160-164.

2. Jordan BD. The clinical spectrum of sport-related traumatic brain injury. *Nat. Rev. Neurol.* 2013; 9(4):222-230.

3. McCrory P, Meeuwisse WH, Aubry M, et al. Consensus statement on concussion in sport: the 4th International Conference on Concussion in Sport held in Zurich, November 2012. *Br. J. Sports Med.* 2013; 47(5):250-258.

4. Fuller GW, Tucker R, Starling L, Falvey E, Douglas M, Raftery M. The performance of the World Rugby Head Injury Assessment Screening Tool: a diagnostic accuracy study. *Sports medicine - open.* 2020; 6(1):2.

5. Fuller CW, Fuller GW, Kemp SP, Raftery M. Evaluation of World Rugby's concussion management process: results from Rugby World Cup 2015. *Br. J. Sports Med.* 2017; 51(1):64-69.

6. Lezak MD. *Neuropsychological assessment*, New York, Oxford University Press; 2012.

7. Sherman EMS, Spreen O, Strauss E. *Compendium of neuropsychological tests : administration, norms, and commentary*, New York, Oxford University Press; 2006.

8. Ben-David BM, Nguyen LL, van Lieshout PH. Stroop effects in persons with traumatic brain injury: selective attention, speed of processing, or color-naming? A meta-analysis. *J. Int. Neuropsychol. Soc.* 2011; 17(2):354-363.

9. Lange RT, Iverson GL, Zakrzewski MJ, Ethel-King PE, Franzen MD. Interpreting the trail making test following traumatic brain injury: comparison of traditional time scores and derived indices. *J. Clin. Exp. Neuropsychol.* 2005; 27(7):897-906.

10. Woods DL, Wyma JM, Herron TJ, Yund EW. The Effects of Repeat Testing, Malingering, and Traumatic Brain Injury on Computerized Measures of Visuospatial Memory Span. *Front. Hum. Neurosci.* 2015; 9:690.

11. Patricios J, Fuller GW, Ellenbogen R, et al. What are the critical elements of sideline screening that can be used to establish the diagnosis of concussion? A systematic review. *Br. J. Sports Med.* 2017; 51(11):888-894.

12. Leiter R. Partington's pathway test. *Psychol Serv Cent Bull.* 1949; 1:9-20.

13. Salthouse TA. What cognitive abilities are involved in trail-making performance? *Intelligence.* 2011; 39(4):222-232.

14. Scarpina F, Tagini S. The Stroop Color and Word Test. *Front. Psychol.* 2017; 8(557).

15. Stroop JR. Studies of interference in serial verbal reactions. *J. Exp. Psychol. Gen.* 1992; 121(1):15-23.

16. Corsi PM. Human memory and the medial temporal region of the brain. Montreal: McGill Univ.; 1973.

17. Kessels RP, van Zandvoort MJ, Postma A, Kappelle LJ, de Haan EH. The Corsi Block-Tapping Task: standardization and normative data. *Appl. Neuropsychol.* 2000; 7(4):252-258.

18. Raftery M, Kemp S, Patricios J, Makdissi M, Decq P. It is time to give concussion an operational definition: a 3-step process to diagnose (or rule out) concussion within 48 h of injury: World Rugby guideline. *Br. J. Sports Med.* 2016; 50(11):642-643.

19. Echemendia RJ, Meeuwisse W, McCrory P, et al. The Sport Concussion Assessment Tool 5th Edition (SCAT5): Background and rationale. *Br. J. Sports Med.* 2017; 51(11):848-850.

20. Collie A, Maruff P, Makdissi M, McCrory P, McStephen M, Darby D. CogSport: reliability and correlation with conventional cognitive tests used in postconcussion medical evaluations. *Clin. J. Sport Med.* 2003; 13(1):28-32.

21. Fukui T, Sugita K, Sato Y, Takeuchi T, Tsukagoshi H. Cognitive functions in subjects with incidental cerebral hyperintensities. *Eur. Neurol.* 1994; 34(5):272-276.

22. Hashimoto R, Meguro K, Lee E, Kasai M, Ishii H, Yamaguchi S. Effect of age and education on the Trail Making Test and determination of normative data for Japanese elderly people: the Tajiri Project. *Psychiatry Clin. Neurosci.* 2006; 60(4):422-428.

23. Armengol CG. Stroop Test in Spanish: Children's Norms. *The Clinical Neuropsychologist.* 2002; 16(1):67-80.

24. Heaton RK, Grant I, Matthews CG. *Comprehensive norms for an expanded Halstead-Reitan battery: demographic corrections, research findings, and clinical applications; with a supplement*

for the Wechsler Adult Intelligence Scale-Revised (WAIS-R), Psychological Assessment Resources; 1991.

25. Brooks A, MacDonald K, Wasylyk N, Hetzel S, McGuine T. NORMATIVE VALUES FOR THE SCAT3 IN ADOLESCENT ATHLETES. *Br. J. Sports Med.* 2014; 48(7):573-574.

26. Fuller GW, Govind O, Tucker R, Raftery M. Sport concussion assessment tool-Third edition normative reference values for professional Rugby Union players. *J. Sci. Med. Sport.* 2018; 21(4):347-351.

27. Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad. Emerg. Med.* 1996; 3(9):895-900.

28. Bigler ED. *Neuropathology of Mild Traumatic Brain Injury: Correlation to Neurocognitive and Neurobehavioral Findings Brain Neurotrauma: Molecular, Neuropsychological, and Rehabilitation Aspects*, Boca Raton FL, 2015 by Taylor & Francis Group, LLC.; 2015.

29. Bohnen N, Jolles J, Twijnstra A. Modification of the stroop color word test improves differentiation between patients with mild head injury and matched controls. *Clin. Neuropsychol.* 1992; 6(2):178-184.

30. Koss E, Ober BA, Delis DC, Friedland RP. The Stroop color-word test: indicator of dementia severity. *Int. J. Neurosci.* 1984; 24(1):53-61.

31. Sisco SM, Slonena E, Okun MS, Bowers D, Price CC. Parkinson's disease and the Stroop color word test: processing speed and interference algorithms. *The Clinical neuropsychologist.* 2016; 30(7):1104-1117.

32. Lansbergen MM, Kenemans JL, van Engeland H. Stroop interference and attention-deficit/hyperactivity disorder: a review and meta-analysis. *Neuropsychology.* 2007; 21(2):251-262.

33. Yengo-Kahn AM, Hale AT, Zalneraitis BH, Zuckerman SL, Sills AK, Solomon GS. The Sport Concussion Assessment Tool: a systematic review. *Neurosurg. Focus.* 2016; 40(4):E6.

34. Soukup VM, Ingram F, Grady JJ, Schiess MC. Trail Making Test: issues in normative data selection. *Appl. Neuropsychol.* 1998; 5(2):65-73.

35. Perianez JA, Rios-Lago M, Rodriguez-Sanchez JM, et al. Trail Making Test in traumatic brain injury, schizophrenia, and normal ageing: sample comparisons and normative data. *Arch. Clin. Neuropsychol.* 2007; 22(4):433-447.

Table 1. Agreement of repeated pre-season baseline cognitive tests evaluated using the Bland-Altman limits of agreement, concordance correlation, and weighted kappa approaches.

	Cognitive test					
	TMT-B time n=216	**TMT-B errors** n=216	**Stroop test time** n=219	**Stroop test score** n=219	**Spatial memory test score** n=226	**Spatial memory test failed trials** n=226
Mean of paired differences (95% CI)	1.2 (0.1-2.4)	-	1.8 (0.8-2.7)	-	12.6 (1.8-23.4)	-
Lower 95% limit of agreement	-15.1	-	-12.0	-	-149.2	-
Upper 95% limit of agreement	17.6	-	15.6	-	174.3	-
Concordance correlation coefficient (95% CI)	0.3 (0.2-0.4)	-	0.2 (0.1-0.3)	-	0.2 (0.1-0.3)	-
Bias correction factor	0.9	-	0.9	-	0.9	-
Pearson's correlation coefficient (95% CI)	0.3 (0.2-0.5)	-	0.2 (0.1-0.4)	-	0.2 (0.1-0.3)	-
Absolute agreement (%, 95% CI)	-	37.0% (30.5-43.5)	-	22.8% (17.2-28.4)	-	49.1% (42.6-55.7)
Weighted kappa (95% CI)	-	0.01 (0.0- 0.1)	-	0.2 (0.1 - 0.3)	-	0.1 (0.0 - 0.2)
Gwets AC1	-	0.3 (0.3- 0.4)	-	0.1 (0.1-0.2)	-	0.3 (0.2-0.4)

Table 2. Performance of cognitive tests for the detection of concussion

N=100	Median value of distribution			AUROC	Normative test threshold	Diagnostic accuracy using Normative cut-point		Additive performance to HIA01 off-field screening tool	
Cognitive test	**Concussed**	**Non-concussed**	**p-Value**	**AUROC (95% CI)**		**Sensitivity (%)**	**Specificity (%)**	**Additional concussions detected**	**Additional false positives**
						(95% CI)			
Trail Making Test B - Time	17.2	17.8	0.8	0.5 (0.4-0.6)	≥27.7	12.1 (3.4-28.2)	86.6 (76.9-93.7)	0	4
Trail Making Test B – Errors	1	1	1.0	0.5 (0.4-0.6)	≥4	15.2 (5.1-31.9)	86.6 (76.0-93.7)	1	6
Stroop test – Time (secs)	21.1	18.4	0.01*	0.7 (0.5-0.8)	>25.5	15.2 (5.1-31.9)	91.0 (81.5-96.6)	2	4
Stroop test- Score	10	9	0.6	0.5 (0.4-0.6)	≤4	12.1 (3.4-28.2)	89.6 (79.7-95.7)	1	5
Spatial memory test – Score	280	270	1.0	0.5 (0.4-0.6)	≤205	24.2 (11.1-42.3)	79.1 (67.4-88.1)	1	8
Spatial memory test – Failed trials	1	1	0.6	0.5 (0.4-0.6)	≥2	33.3 (18.0-51.8)	65.7 (53.1-76.8)	2	13

*Significant at α=0.05

Figure legends

Figure 1. Flow diagram showing derivation of study sample.