RESEARCH ARTICLE

REVISED One-year test-retest reliability of ten vision tests in Canadian athletes [version 5; peer review: 2 approved]

Mehdi Aloosh[1,2], Suzanne Leclerc[3], Stephanie Long[4], Guowei Zhong[4], James M. Brophy[5], Tibor Schuster[4], Russell Steele[6], Ian Shrier[4,7]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[2]Department of Health Research Methods, Evidence, and Impact, Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Canada
[3]Institut National du Sport du Quebec, Montreal, Canada
[4]Department of Family Medicine, McGill University, Montreal, Canada
[5]Faculty of Medicine, McGill University, Montreal, Canada
[6]Department of Mathematics and Statistics, McGill University, Montreal, Canada
[7]Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, Canada

v5 **First published:** 09 Jul 2019, **8**:1032
https://doi.org/10.12688/f1000research.19587.1

Second version: 31 Mar 2020, **8**:1032
https://doi.org/10.12688/f1000research.19587.2

Third version: 08 Jun 2020, **8**:1032
https://doi.org/10.12688/f1000research.19587.3

Fourth version: 26 Aug 2020, **8**:1032
https://doi.org/10.12688/f1000research.19587.4

Latest published: 09 Sep 2020, **8**:1032
https://doi.org/10.12688/f1000research.19587.5

Abstract

Background: Vision tests are used in concussion management and baseline testing. Concussions, however, often occur months after baseline testing and reliability studies generally examine intervals limited to days or one week. Our objective was to determine the one-year test-retest reliability of these tests.
Methods: We assessed one-year test-retest reliability of ten vision tests in elite Canadian athletes followed by the Institut National du Sport du Quebec. We included athletes who completed two baseline (preseason) annual evaluations by one clinician within 365±30 days. We excluded athletes with any concussion or vision training in between the annual evaluations or presented with any factor that is believed to affect the tests (e.g. migraines). Data were collected from clinical charts. We evaluated test-retest reliability using Intraclass Correlation Coefficient (ICC) and 95% limits of agreement (LoA).
Results: We examined nine female and seven male athletes with a mean age of 22.7 (SD 4.5) years. Among the vision tests, we observed excellent test-retest reliability in Positive Fusional Vergence at 30cm (ICC=0.93) but this dropped to 0.53 when an outlier was excluded in a sensitivity analysis. There was good to moderate reliability in Negative Fusional Vergence at 30cm (ICC=0.78), Phoria at 30cm (ICC=0.68), Near Point of Convergence break (ICC=0.65) and Saccades (ICC=0.61). The ICC for Positive Fusional Vergence at 3m (ICC=0.56) also decreased to

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 5 (revision) 09 Sep 2020		✓ report
version 4 (revision) 26 Aug 2020	✓ report	? report
version 3 (revision) 08 Jun 2020		? report
version 2 (revision) 31 Mar 2020		
version 1 09 Jul 2019	✓ report	

1. **M Nadir Haider**, State University of New York at Buffalo, Buffalo, USA

0.45 after removing two outliers. We found poor reliability in Near Point of Convergence (ICC=0.47), Gross Stereoscopic Acuity (ICC=0.03) and Negative Fusional Vergence at 3m (ICC=0.0). ICC for Phoria at 3m was not appropriate because scores were identical in 14/16 athletes. 95% LoA of the majority of tests were ±40% to ±90%.
Conclusions: Five tests had good to moderate one-year test-retest reliability. The remaining tests had poor reliability. The tests would therefore be useful only if concussion has a moderate-large effect on scores.

Keywords
concussion, vision tests, binocular, saccades, reliability

2. **Dillon Richards**, University of Western Ontario, London, Canada

James P Dickey, University of Western Ontario, London, Canada

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Ian Shrier (ian.shrier@mcgill.ca)

Author roles: Aloosh M: Writing – Original Draft Preparation, Writing – Review & Editing; **Leclerc S**: Conceptualization, Funding Acquisition, Resources, Writing – Review & Editing; **Long S**: Data Curation, Writing – Review & Editing; **Zhong G**: Data Curation, Writing – Review & Editing; **Brophy JM**: Validation, Writing – Review & Editing; **Schuster T**: Formal Analysis; **Steele R**: Formal Analysis, Writing – Review & Editing; **Shrier I**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project was funded by government sources (MITACS [IT08159] and MEDTEQ [G245120], a non-profit organization (Institut National du Sport du Quebec) and private industry (APEXk Inc, and Varitron Inc.)
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Aloosh M, Leclerc S, Long S *et al.* **One-year test-retest reliability of ten vision tests in Canadian athletes [version 5; peer review: 2 approved]** F1000Research 2020, **8**:1032 https://doi.org/10.12688/f1000research.19587.5

First published: 09 Jul 2019, **8**:1032 https://doi.org/10.12688/f1000research.19587.1

REVISED **Amendments from Version 4**

The reviewers insisted that we consider our sensitivity analyses excluding outliers as the primary analysis. Although we do not believe this is the optimal approach, we made this change in the current version. There are some other small edits about some values from this athlete population being outside the normative data in the general population that we obtained from the literature.

Any further responses from the reviewers can be found at the end of the article

Introduction

Concussion, a form of mild traumatic brain injury is a growing public health concern[1]. Estimates suggest up to 3.8 million sport-related concussions occur annually in the United States, with 50% going unreported[2]. United States emergency department visits for sports-related traumatic brain injuries have increased 60% over 2001–2009[3]. Concussions can be associated with headaches, dizziness, visual disturbances, and other symptoms that can negatively affect performance in sport, school, and work and negatively impact quality of life[2,4,5].

Diagnosis of concussion and decisions to return-to-play are based on symptoms, signs, physical examination and special tests[6]. Previous research has shown an association between concussion and eye movement[1]. Concussion may therefore affect multiple aspects of vision, including saccades, pursuit, convergence, accommodation, and vestibulo-ocular reflex[7]. Some studies reported 50% to 90% incidence of visual symptoms, such as blurred vision and diplopia in individuals with concussion[8]. Therefore, vision testing may be helpful in the assessment and management of patients with concussion.

Each vision test measures a function that is linked to a particular brain structure or pathway. Vision tests are noninvasive tests with rapid administration and scoring. Understanding test variability, independent of changes in pathology or recovery (i.e. reliability), is required to assess their clinical utility. However, only a limited number of reliability studies have assessed binocular vision tests and saccades[9–20]. In addition, these reliability studies measured a specific aspect of the vision. These studies are not uniform in their method and they are diverse in their population.

Previous investigations of the test-retest reliability of these vision tests have used short test-retest time intervals ranging from 0 to approximately 57 days[9–20], except for one test of saccades[21]. For test-retest reliability to be useful in clinical management (e.g. return-to-play), the time intervals must reflect the time frame in which they would be used[22]. The previous studies have provided information on the usefulness of these tests when following improvement or deterioration of patients over short periods of time. However, concussions usually occur several months and up to one year after annual baseline testing, and not as 0 days to 57 days as in the previous studies. Therefore, we examined one-year test-retest reliability of ten vision tests in Canadian athletes over one year period of time.

Methods

Participants

The study population included athletes over 16 years of age followed by the Institut National du Sport du Quebec (INSQ) in Canada from 2015–2018. Many of these athletes had a yearly examination done by a sports medicine physician and vision tests done by a clinician trained in orthoptic testing.

We only included athletes who had completed two baseline (preseason) annual evaluations within a 365-day (± 30 days) time period. We excluded athletes who suffered a concussion in between annual evaluations or had received preventive orthoptic training between the baseline measures. We also excluded athletes with a history of strabismus or treated strabismus, or were medically treated for depression, anxiety or psychiatric conditions that may affect binocular vision and saccades. Data were collected from electronic medical charts of one clinician trained in orthoptic measures and one sports medicine physician.

Measures

At the beginning of each season, athletes underwent baseline testing of ten vision tests by a single orthoptic-trained clinician (industry partner). The vision tests were Gross Stereoscopic Acuity, Near Point of Convergence (NPC), Near Point of Convergence break (NPCb), near (30cm) and far (3m) Positive Fusional Vergence, near (30cm) and far (3m) Negative Fusional Vergence, near (30cm) and far (3m) Phoria, and Saccades.

A detailed description of each test including the procedures of each test and the theoretical range of scores is provided in Table 1. We will briefly describe each vision test here. We used a horizontal prism bar with the base-out for Positive Fusional Vergence and base-in for Negative Fusional Vergence, at both 30cm and 3m[10]. Phoria was measured at 30cm and 3m using the prism and alternate cover test using the procedures described by the Pediatric Eye Disease Investigator Group[23]. To perform NPC and NPCb, we followed the Maples *et al.*, protocol[13]. We measured Gross Stereoscopic Acuity with the Randot Stereotest (Stereo Optical Co., Inc., Chicago, IL) according to the manufacturer's instructions[24]. Evaluation of Saccades was done using the test procedures developed by the orthoptic-trained clinician. Participants assumed a tandem stance an arm's length away from a screen attempting to fixate on appearing and disappearing lights on the screen, while trying to keep their head still. Light flashes appeared at a rate of 100 per minute for two minutes. This test was scored by the clinician based on quality (bad, medium, good), synchronization (bad, medium, good), and saccadic corrections (many, few, none). These three components were then combined into an overall percentage saccade score, based on an unpublished proprietary algorithm developed by the clinician who performed the testing.

Table 1. Detailed description of the ten vision tests.

Positive Fusional Vergence	This test examines how well a participant can adapt to challenges in fixating light on their retina at near distance (30cm) and far distance (3m), measured in prism diopters. The seated participant fixates on a fixed target at the appropriate distance. The clinician begins by using the weakest prism strength (base-out) which forces the participant to converge their eyes to maintain fixation. The strength of the prism is increased until the participant can no longer maintain a single image. The score of each test (30cm and 3m) is the strength of the prism in which the participant maintained binocular vision, with higher scores representing better function. The range of normative data for Positive Fusional Vergence at near fixation is 35 to 40 prism diopters, and the range at far fixation is 16 to 20 prism diopters[25–27].
Negative Fusional Vergence	This is the same test as Positive Fusional Vergence except the horizontal prism bar is positioned base-in, forcing the participant to diverge their eyes to maintain fixation on a fixed object positioned at near (30cm) and far (3m), measured in prism diopters. The clinician incrementally increases the strength of the prism until the participant is no longer able to maintain a single image. The score of each test is the strength of the prism in which the participant maintained binocular vision, with higher scores representing better function. The range of normative data for Negative Fusional Vergence at near fixation is 12 to 16 prism diopters, and the range at far fixation is 6 to 8 prism diopters[25–27].
Phoria	We evaluated the natural deviation of the eyes (heterophoria), in prism diopters, with the prism and alternate cover test using a target placed at (1) 3m from the participant (far vision), and (2) 30cm from the participant (near vision). While the seated participant was fixating on the target, the clinician covered and uncovered each of the participant's eyes to trigger movements while using a prism bar (base-out if the eye moves outward, base-in if the eye moves inward) to cancel these movements. The prism power was progressively increased until no shift in the eyes was seen. The score of the test was the rating of the prism that canceled the eye movements, with lower scores representing less Phoria. We were unable to find normative data for this test.
Near Point of Convergence (NPC)	NPC assesses the ability to symmetrically converge, and is sometimes referred to as "motor punctum proximum"[26], in cm. The seated participant fixates on a near target 30cm away. The target is gradually moved towards their eyes as they attempt to maintain fixation. NPC is reached when one or both eyes can no longer maintain fixation on the target, which is identified as when one eye diverges outwards. The score of the test is the distance (cm) between the bridge of the nose and the distance of the target at the closest point at which the individual could maintain balanced oculomotor synergy between both eyes. Lower scores indicate better NPC. Normative data in older textbooks report average NPC values for healthy adults between 6 to 8 cm[28], but a more recent study suggested 5 cm should be considered the upper limit of normal values[29].
Near Point of Convergence break (NPCb)	This test is conducted using the same methods as NPC, but the test ends when the participant has double vision due to the inability of the eyes to converge. The score of the test is the distance between the bridge of the nose and the point (in cm) where double vision occurs, where a lower score indicates better NPCb. Normative data for elementary school children with normal vision suggested a mean of 3.3 cm, with a range of 1.0 to 13.7 cm[30]; however, data on adults with normal vision suggest a breakpoint of approximately 5.0 to 7.5 cm[31].
Gross Stereoscopic Acuity	We tested the ability to perceive depth with the Randot® Stereotest (Stereo Optical Co., Inc., Chicago, IL), in arc seconds. Seated participants wearing polarized glasses were asked to hold the testing booklet 16 inches from their face. Participants were then presented images formed of dots that are displaced in relation to each other. The test steadily increased in difficulty by reducing the level of disparity between dots, beginning at 400 arc seconds (lowest possible score) and ending at 20 arc seconds (highest possible score). A participant's score was the arc seconds corresponding to the smallest disparity at which the participant identified the raised (i.e. stereoscopic) image. Normative data suggest the average score for an adult is 40 arc seconds[32,33].
Saccades	This test examines the eye's ability to perform saccadic movements, which are rapid eye movements that abruptly alter the point of fixation. In our clinician's version of this test, participants assume a tandem stance (heel-to-toe with dominant foot in the back) standing an arm's length away from the screen. Lights appear and disappear in different locations on the screen at a rate of 100 flashes per minute, for a total of two minutes. The participant is instructed to keep their head still and only move their eyes to fixate on the appearing lights. The clinician observes the eyes for quality and synchronization (rated: bad, medium, good) and saccadic correction (rated: many corrections, few corrections, no corrections). The three sub-scores were combined into an overall percentage score according to a proprietary algorithm developed by the clinician (industry partner) who performed the testing. There are no normative data for this version of the test because the score is based on a proprietary algorithm.

Analysis

We report the mean (SD) for continuous variables at baseline. We evaluated test-retest reliability using Intraclass Correlation Coefficient (ICC)[34] and 95% limits of agreement (LoA)[35]. We considered ICC of ≤0.5 as poor, 0.51–0.74 as moderate, 0.75–0.89 as good, and ≥0.90 as excellent reliability[36]. We report the LoA in the raw units of the scale used by clinicians. To compare LoA across tests, we also standardized the scores and reported them as percent differences, [(T1- T2)/mean(T1&T2)]*100[35,37]. Additionally, we summarized LoA graphically with Bland-Altman plots for each vision test using the standardized score for the y-axis to provide an overview of all vision tests. The raw scale measures are provided in parentheses to provide clinicians with information for individual patient assessment. Finally, we conducted a sensitivity analysis for the vision tests by excluding outliers that may have augmented the ICC results. We defined an outlier as a data point that was 1.5 interquartile ranges below the first quartile or above the third quartile.

Due to the limited sample size (n=16) and to avoid being overly conservative in our evaluation, we followed the practical solution for addressing multiple testing proposed by Saville, the unrestricted least significant difference procedure (or multiple t-test)[38]. Formal multiplicity correction of confidence levels was not performed but we thoroughly reported all statistical assessments enabling an informal type-I error assessment by the reader. The data were analyzed using R statistical software 3.4.3[39]. This study was approved by the McGill University Faculty of Medicine Institutional Review Board.

Results

Of the 199 athletes measured for the vision tests, only 16 individuals met our inclusion criteria (Figure 1). There were nine female and seven male athletes with a mean age of 22.7 (4.5) years at the baseline (preseason) measurement. Participants were athletes of water polo (n=6) and short-track speed skating (n=10). A second measurement was conducted between 335 and 372 days (mean of 356.4 (17.3) days) after the initial baseline.

The range of scores observed for each vision test can be found in each of the reliability figures (Figure 2–Figure 4)[40]. Our analysis suggested one-year test-retest reliabilities ranging from poor to excellent among the ten vision tests. Including all the data, we observed excellent one-year test-retest reliability in Positive Fusional Vergence at 30cm with ICC of 0.93 (Figure 2). In this test, 4 out of 16 pairs of measurements were identical after 1 year. The range of measurements was between 14 and 45 diopters with one outlier at 90 diopters. LoA of the test was ±41.9%. Given the very high ICC and the presence of an outlier that greatly increased the range of the values for the measure (known to increase ICC), we repeated the analysis excluding the outlier. This decreased the ICC from 0.93 to 0.53, and increased the LoA to ±43.5%. One of the reviewers for this paper has insisted that the analysis without the outlier be considered the primary analysis.

Five tests showed good to moderate one-year test-retest reliability (Figure 3), including Negative Fusional Vergence at 30cm (ICC=0.78, LoA=41.2%), Phoria at 30cm (ICC=0.68, LoA=119.2%), NPCb (ICC=0.65, LoA=49.4%), Positive Fusional Vergence at 3m (ICC=0.56, LoA=60.2%), and Saccades (ICC=0.61, LoA=24.3%). There were two outliers for Positive Fusional Vergence at 3m (one participant on both measures and one participant on only one measure). When we removed both of these outliers, the ICC dropped from 0.56 to 0.45 and the 95% LoA decreased from 60.2% to 41.4%. In both of these cases, the two scores from the outlier were quite different. Although one might anticipate that the ICC would increase by removing such outliers, the ICC actually decreased because the range of values for the measure decreased substantially. As above, one of the reviewers for this paper insisted that the analysis without the outliers be considered the primary analysis.

Three of the remaining four tests showed poor one-year test-retest reliability (Figure 4). These include NPC (ICC=0.47, LoA=73.9%), Gross Stereoscopic Acuity (ICC=0.03, LoA=92.5%) and Negative Fusional Vergence at 3m (ICC=0.0, LoA=48.4%). For Phoria at 3m, 14/16 athletes had identical scores on the two measures. In this context, the ICC and LoA were not appropriate measures of reliability and are not presented.

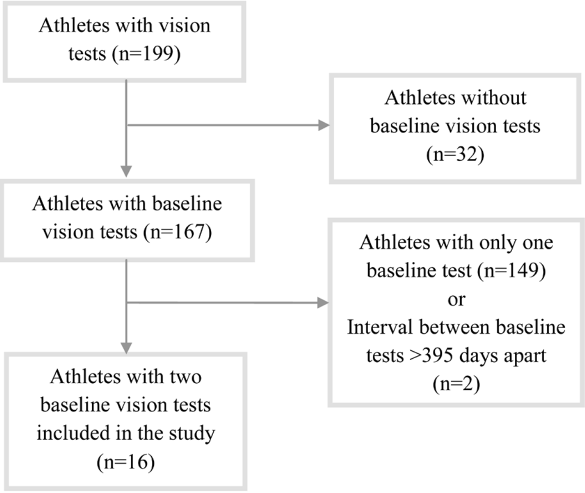
Athletes with vision tests (n=199)
Athletes without baseline vision tests (n=32)
Athletes with baseline vision tests (n=167)
Athletes with only one baseline test (n=149) or Interval between baseline tests >395 days apart (n=2)
Athletes with two baseline vision tests included in the study (n=16)

Figure 1. Patient flow diagram.

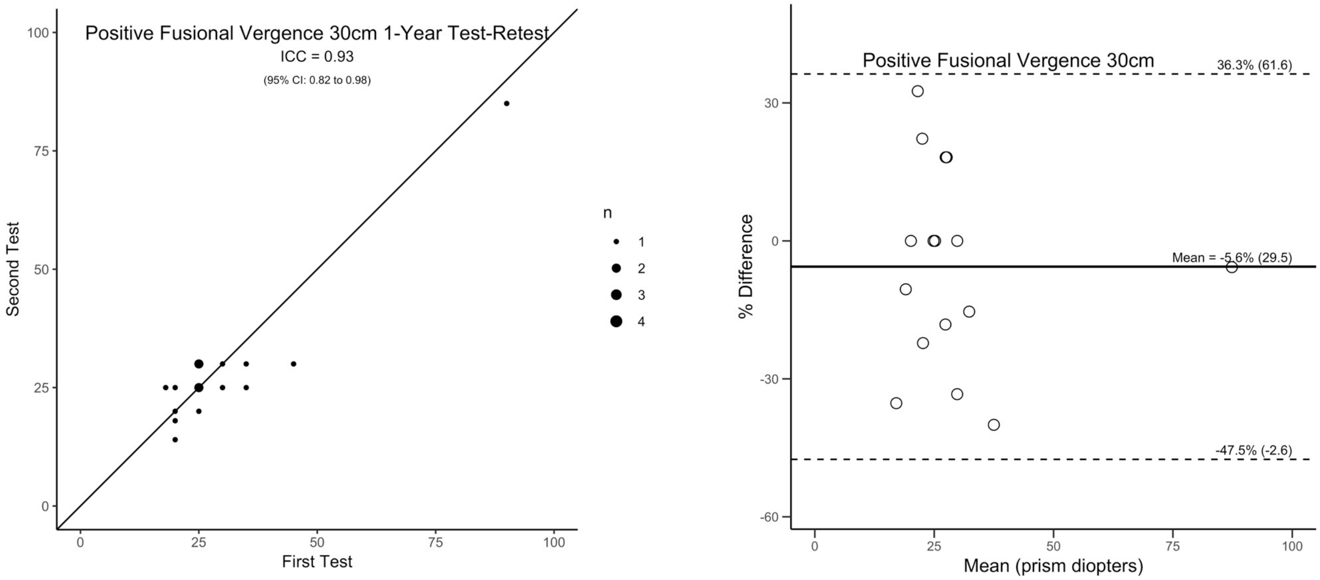

Figure 2. Vision test with excellent one-year test-retest reliability. (**A**) Scatter plot of test-retest reliability for Positive Fusional Vergence at 30cm. Identity line represents perfect agreement between the test-retest values; ICC refers to the Intraclass correlation coefficient and 95%CI refers to the 95% Confidence Interval. "n (1,2,3,4)" refers to the number of participants represented by each dot when scores exactly overlapped. (**B**) Bland-Altman plot with the mean of the test-retest on the x-axis and the difference between test-retest on the y-axis. Solid line represents the bias and dotted lines represent the 95% LoA. The y-axis represents a standardized LoA using percentage difference on the plot to allow one to compare the different tests to each other. The LoA in the units of measure, which are familiar to clinicians, are provided in the parentheses. When the analysis was repeated excluding the outlier to the far right, the ICC decreased to 0.53 and the 95% LoA increased to 43.5%.

Discussion

We found that the one-year test-retest reliability for 10 vision tests in young elite athletes ranged from moderate to poor after accounting for outliers. The majority of the vision tests had standardized 95% LoA in the range of 40–90%, which indicates that repeated scores of an individual over time may vary by 40–90% of the mean score even without any actual change in vision function.

There are a limited number of test–retest reliability studies on non-vision neurocognitive tests over a one year period in teenage athletes. For instance, the ICC for different components of Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT), a computerized brain injury measurement tool, ranges from 0.50 to 0.82[41]. However, we could not find any research examining the stability of the vision tests over a one year period, in athlete or non-athlete populations except for one test of saccades that was very different from the test used in this study[21]. It is important that test-retest reliabilities fall within a range needed for clinical interpretation of concussion assessment and for discussion about return-to-play. In the context of comparing results after a concussion to annual baseline tests conducted in the pre-season, the time-frame for reliability comparisons should be up to one year[22].

Although there are no long-term reliability studies on the ten vision tests evaluated in this study, a number of studies have reported short term test-retest reliability of individual tests using various methods among various groups of individuals, including children and healthy adults[9–20]. Using NPC as a general example, one study reported excellent immediate test-retest reliability in concussed athletes (ages 9–24) (ICC = 0.95 to 0.98)[12]. A separate study using a 2–3 day test-retest protocol found the ICC = 0.65 for NPC in healthy individuals (calculated in Rouse *et al.*, 2002[16] for data from reference[15]), and a third study reported one week test-retest ICC = 0.89 and 0.92 for NPCb in healthy school children[16].

We recently examined one-week test-retest reliability of the same ten vision tests with the same methods and same age-range as this current study in 20 young non-athletes. We found one-week test-retest reliability ranging from poor (ICC = 0.34) to good (ICC = 0.88), with five out of ten tests showing moderate reliability (ICCs = 0.54 to 0.69)[17]. This suggests that these vision tests can only be useful if a concussion has a moderate to large effect on scores. Overall, the ICCs in the current study were generally smaller than those reported in our one-week study, suggesting increased temporal variability. Unexpectedly, the 95% LoA for one-year test-retest was smaller or equal to the 95% LoA of the one-week test-retest for all vision tests except NPC (±73.9 vs. ±57.9) and Gross Stereoscopic Acuity (±92.5 vs. ±55). In addition, in both the one-week and one-year intervals, almost all individuals had the same value in Phoria 3m, which leads to uninformative LoA.

In one-year test-retest, Positive Fusional Vergence showed excellent reliability at 30cm (ICC=0.93) and moderate at 3m (ICC=0.56), initially. Our results at 30cm were significantly better than those of another study examining test-retest reliability of Positive Fusional Vergence at 30cm in children (ICCs of 0.53–0.59)[16]. Perhaps more importantly, our results were also better than the one-week test-retest reliability conducted by the

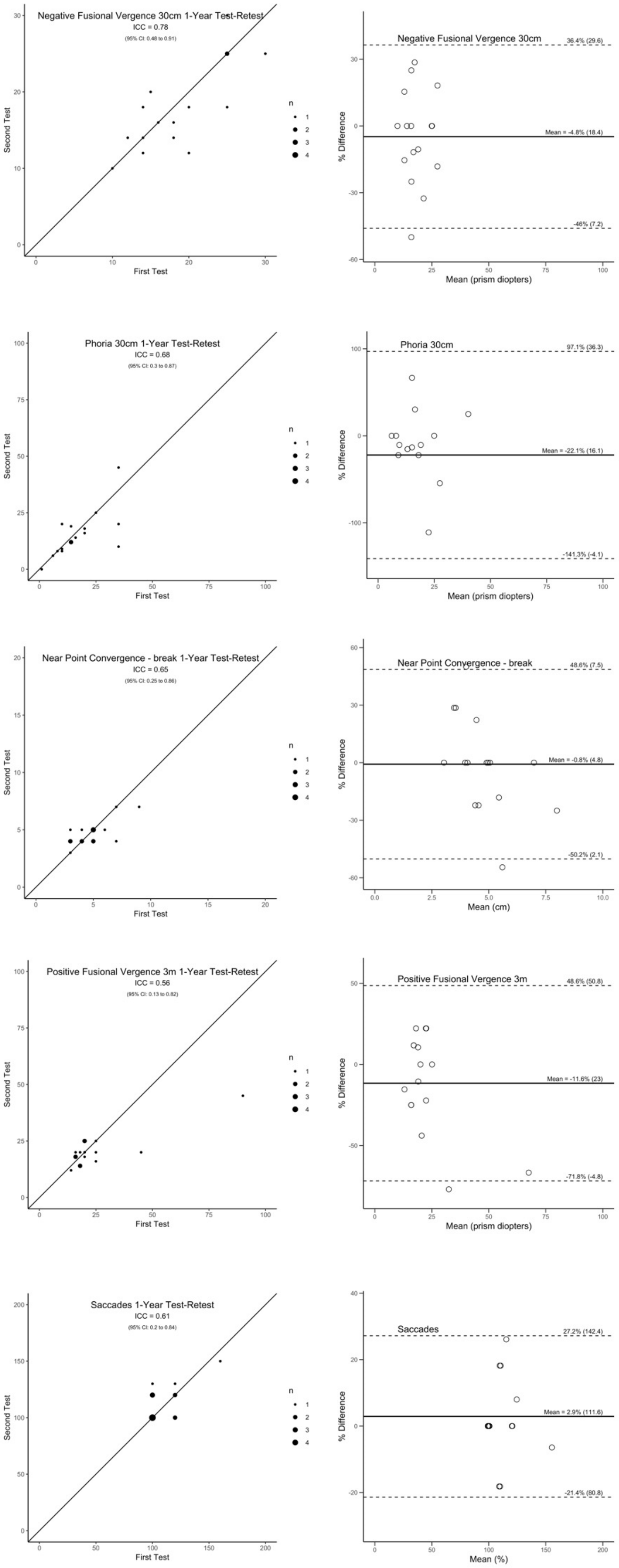

Figure 3. Vision tests with good to moderate one-year test-retest reliability. (**A**) Scatter plot of test-retest for Negative Fusional Vergence at 30cm, Phoria at 30cm, Near Point of Convergence break (NPCb), Positive Fusional Vergence at 3m, and Saccades. (**B**) Bland-Altman plot related to each test. See Figure 2 for explanation of abbreviations and scales. When the analysis for Positive Fusional Vergence at 3m was repeated excluding the two outliers, the ICC decreased to 0.45 and the 95% LoA decreased to 41.4%.

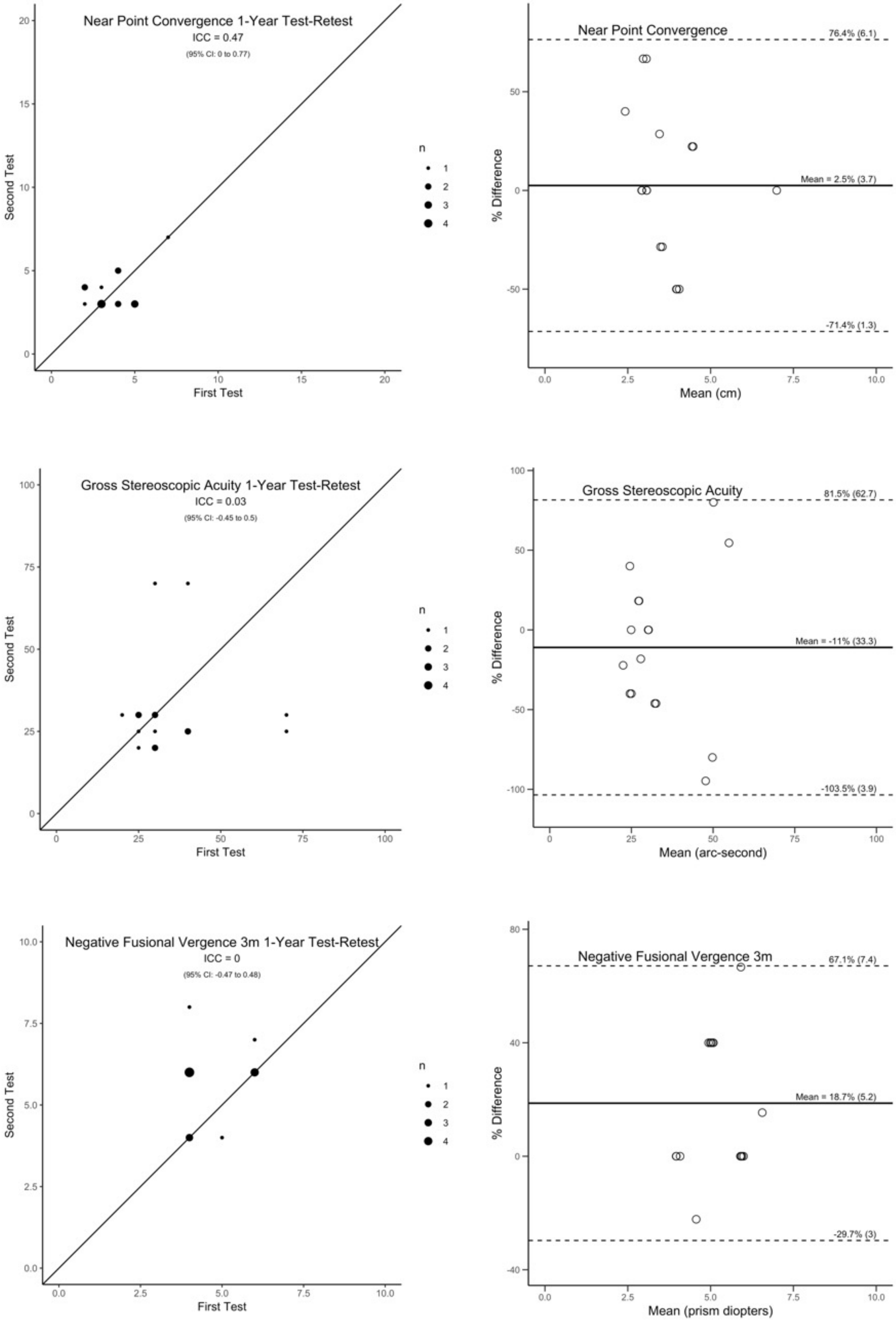

Figure 4. Vision tests with poor one-year test-retest reliability. (**A**) Scatter plots of test-retest for near point of convergence (NPC), Gross Stereoscopic Acuity, and Negative Fusional Vergence at 3m. (**B**) Bland-Altman plots related to each test. See Figure 2 for explanation of abbreviations and scales.

same clinician with the same methods in our previous prospective research study (ICC=0.54 and 0.49, respectively)[17]. It is difficult to understand how test-retest reliability over one year could be better than test-retest reliability over one week. When we explored the data further, we noticed one outlier that greatly increased the range of values for Positive Fusional Vergence at 30cm (Figure 2) and Positive Fusional Vergence at 3m (Figure 3). Increasing the range of values is known to increase the ICC. This is because ICC is based on the results of an analysis of variance which separates the error into variability between individuals (range of values along x or y axes) and variability within an individual. Therefore, if variability between persons increases, indicated by a larger range of values, ICC will increase. When we removed the outlier for Positive Fusional Vergence at 30cm, the ICC dropped to 0.53, which is similar to the value found for the one-week test-retest reliability (ICC=0.54); the LoA increased to 43.5%. When we removed the two outliers from Positive Fusional Vergence at 3m, the ICC decreased to 0.45 and LoA decreased to 41.4%. Note that the outliers for this measure had large differences between the two test scores, and removing such data points would normally be expected to increase the ICC (Figure 3). The finding that the ICC decreased indicates that as expected, if the range of values among the populations is similar, the one-year test-retest reliability for Positive Fusional Vergence at both 30cm and 3m is likely less than the one-week test-retest reliability.

In addition to Positive Fusional Vergence, two other tests also had higher ICC at one year (Negative Fusional Vergence 30cm: 0.78 vs 0.66) and Saccades (0.61 vs 0.34) but there were no apparent outliers and the range of values were similar in the two studies. Aside from outliers, there are other theoretical reasons that might explain why ICC is better at one-year than at one-week. First, it is possible that the non-athletes in our one-week test-retest study had less motivation to perform well on the repeat tests. If true, their scores would be less than the motivated athletes performing during the one-year test-retest. Second, there is a potential learning effect in retest measurements that could affect results. A learning effect, however, is unlikely in our study because the athletes were tested only twice, with a one-year interval between tests. Third, the one-week study was a prospective research study where the clinician performing the test was blinded. Our current results are based on clinical charts where the clinician had access to the previous results which might artificially increase the reliability of the test. Fourth, the increased ICC could have occurred simply by chance because of sampling variation.

Our measurements of Phoria at 30cm had moderate reliability for near (ICC=0.68) consistent with our one-week retest reliability study (ICC=0.69)[17]. Other studies in adults and children with strabismus[42] or esotropia[23] have not reported ICC. Therefore, comparing between studies is not possible. Moreover, our analytical methods differed slightly from those studies. We evaluated all angles of deviation together, and other authors analyzed smaller (2–20 Prism Diopter) or larger (>20 Prism Diopter) angles of strabismus separately because of different prism increments measured[42]. For Phoria at 3m, we found that the ICC and LoA were not appropriate measures of reliability because most of the population reported identical scores of zero for both measurements. One may consider that if we had a wider range of scores, ICC might provide meaningful information.

One-year test-retest reliability of NPC and NPCb (0.47 and 0.65, respectively) were similar to the results in our one-week reliability study (0.54 and 0.64, respectively)[17]. Brozek *et al*. found a similar ICC of 0.65 for NPC in healthy adults (calculated in Rouse *et al*., 2002[16] for data from Brozek *et al*., 1948[15]). However, Giffard *et al.* reported a one-week ICC = 0.84 in patients for NPC with neck pain[18] and Rouse *et al.* reported excellent one-week reliability for NPCb in school children (ICC=0.89 and 0.92 for two different examiners)[16]. The discrepancies in results are most likely due to differences in testing procedures. For instance, we used the Maples method[13] which is a non-accommodative test. Rouse *et al.*[16] used an accommodative target with Astron International Accommodative Rule and Giffard *et al.*[18] used the RAF rule[28].

Our one-year test-retest results for Gross Stereoscopic Acuity in young athletes showed poor reliability (ICC=0.03; 95% LoA= ±92.5%) even though our previous one-week test-retest results reported good reliability in non-athlete young adults (ICC=0.86; 95% LoA = ± 54%)[17] and another study using Titmus stereo fly and Frisby stereo tests in pre-school children revealed an excellent one-week reliability (ICC=1.0)[19]. In addition, another study reported that 82.0% of their participants had identical results at test and retest taken on the same day in 100 healthy adult and children[11]. With a one-year ICC of 0.03 and LoA of 92.5%, Gross Stereoscopic Acuity cannot be considered a reliable test to assess the vision function over one year, although it may still be appropriate for use in shorter time intervals, such as one week[11,17,19].

Finally, our clinician's test of Saccades showed moderate reliability (ICC=0.61) with the smallest LoA (in percentage) of other tests, similar to the one-week study[17]. These results are similar to other findings in healthy adults over a two-month period (ICC=0.59)[20]. With a moderate reliability and the smallest LoA amongst the other vision tests, the results of the test of Saccades could be considered stable over a one year period assessing athletes.

In this study, four vision tests (Negative Fusional Vergence at 30cm, Phoria at 30cm, Saccades and NPCb) had moderate one-year test-retest reliability. The one test with identical scores in 14/16 athletes was Phoria at 3m. Therefore we cannot comment on the reliability of this test. This level of reliability would be useful in conditions where the concussion leads to a moderate change in vision function. The remaining five vision tests, including Positive Fusional Vergence at 30cm and 3m, NPC, Negative Fusional Vergence at 3m, and Gross Stereoscopic Acuity may be useful to detect the effect of concussion with a large change on vision function. Further studies are therefore required to assess the effect of concussion on vision test scores of the five vision tests. If it can be shown that the concussion has moderate to large effect on the test scores then these vision tests may still be useful clinically.

Strengths and limitations

Several studies have previously evaluated the inter-rater reliability of some vision tests[23,42]. However, inter-rater reliability is less important in the context of clinical care when patients are followed by one clinician over time. Our study evaluated the test-retest reliability of the ten vision tests over an interval that allows for the normal variation over time expected in clinical practice between baseline measures and subsequent concussions. The ICC represents how much of variability in scores is due to differences between subjects. For instance, the ICC of 0.78 for near Negative Fusional Vergence at 30cm suggests that 78% of the variability in the measurements was due to differences between participants, and 22% was due to normal variations within the measurement. Furthermore, the 95% LoA for each test in our study provides the magnitude of the normal variation that can be expected with repeated measurements. Differences in test results between baseline and diagnosis of a concussion likely represent a true signal of a change in vision function within the patient if these differences are larger than the noise (LoA). In addition, we conducted sensitivity analysis to evaluate the effect of outliers. This analysis suggested that our initial ICC results may have been artificially high for two tests. (Positive Fusional Vergence at 30cm and 3m). Finally, the results of the test of Saccades in this study are based on the unpublished proprietary algorithm developed by the clinician. This limits its applicability for other clinicians.

This is a historical cohort observational study, a study design which has inherent limitations. The data provided were not always as precise as one might expect (e.g. near point convergence measured to the nearest cm). Some data in these athletes appear to be outside the normative range of data previously described for the general population. Because the data were obtained as part of clinical practice, the clinician had access to the results of the first test when conducting the repeat test one year later. The lack of blinding may result in higher agreement between the two tests compared to our blinded one-week research study. However, clinicians are not blinded during normal clinical practice, and therefore the results of this study would represent an expected level of agreement in that context, even if some of the agreement is due to bias. In addition, the sample size was relatively small and composed of healthy athletes, which will limit the generalizability of these findings to other populations. Although we started with a pool of 199 athletes, many athletes were excluded because they only had one baseline test, a concussion occurred in between the two baseline tests, or the second baseline test occurred outside the testing window of 365±30 days. Despite starting with athletes from many sports, only athletes from water polo and short-track speed skating met our eligibility criteria. It is unclear if subconcussion impacts affect neurological function in general[43]. If subconcussion impacts were common in these sports and affected vision testing, we should have seen a systematic decrease in vision capacity between the two tests; this was not observed. Further, if it were present, the effect would be considered part of the "noise" clinicians have to consider when comparing the results from post-concussion and baseline tests. With an effective sample size of 16, the anticipated precision of ICC estimates was +/- 0.25 and the study had 80% power to detect ICC values >= 0.6 and more than 90% power to detect ICC values >=0.7 i.e. rejection of the null hypothesis (Table 1a in 44). Note that a total of >60 individuals were required to exclude ICC values <=0.5 with 80% power and an anticipated true ICC>0.7 (Table 2b in 44).

Conclusion

We found that five out of the ten vision tests (Negative Fusional Vergence at 30cm, Phoria at 30cm, NPCb, Positive Fusional Vergence at 30cm, and Saccades) had good to moderate one-year test-retest reliability. This level of reliability is useful in conditions which produce a moderate change in vision function. The remaining five vision tests may be useful in detecting large effects on vision function. If further studies suggest that the effect of concussion on test scores is moderate to large, these vision tests may still be useful clinically.

Data availability

Open Science Framework: Vision Tests in Concussion. https://doi.org/10.17605/OSF.IO/VB4W8[40]

Demographic data are not available. With only 9 males and 7 females from our clinical source, any demographic information would immediately allow some participants to be identified and therefore this information cannot be shared in order to preserve participant confidentiality.

Acknowledgments

We would like to thank Isabel Pereira for her help throughout the course of this work. In addition, we would like to thank David Tinjust, from Apexk for examining the athletes.

References

1. Sussman ES, Ho AL, Pendharkar AV, *et al.*: **Clinical evaluation of concussion: The evolving role of oculomotor assessments.** *Neurosurg Focus.* 2016; **40**(4): E7.
PubMed Abstract | Publisher Full Text

2. Langlois JA, Rutland-Brown W, Wald MM: **The epidemiology and impact of traumatic brain injury: a brief overview.** *J Head Trauma Rehabil.* 2006; **21**(5): 375–8.
PubMed Abstract | Publisher Full Text

3. Centers for disease control and prevention: **Nonfatal traumatic brain injuries related to sports and recreation activities among persons aged ≤19 years--United States, 2001-2009.** *MMWR Morb Mortal Wkly Rep.* 2011; **60**(39): 1337–42.
PubMed Abstract

4. Dikmen S, Machamer J, Fann JR, *et al.*: **Rates of symptom reporting following traumatic brain injury.** *J Int Neuropsychol Soc.* 2010; **16**(3): 401–11.
PubMed Abstract | Publisher Full Text

5. 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): 250–8.
PubMed Abstract | Publisher Full Text

6. McCrory P, Meeuwisse W, Dvořák J, *et al.*: **Consensus statement on concussion in sport-the 5th international conference on concussion in sport held in Berlin, October 2016.** *Br J Sports Med.* 2017; **51**(11): 838–47.
PubMed Abstract | Publisher Full Text

7. Ventura RE, Balcer LJ, Galetta SL: **The neuro-ophthalmology of head trauma.** *Lancet Neurol.* 2014; **13**(10): 1006–16.
PubMed Abstract | Publisher Full Text

8. Talavage TM, Nauman EA, Breedlove EL, *et al.*: **Functionally-detected cognitive impairment in high school football players without clinically-diagnosed concussion.** *J Neurotrauma.* 2014; **31**(4): 327–38.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Oberlander TJ, Olson BL, Weidauer L: **Test-retest reliability of the king-devick test in an adolescent population.** *J Athl Train.* 2017; **52**(5): 439–45.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Goss DA, Becker E: **Comparison of near fusional vergence ranges with rotary prisms and with prism bars.** *Optometry.* 2011; **82**(2): 104–7.
PubMed Abstract | Publisher Full Text

11. Wang J, Hatt SR, O'Connor AR, *et al.*: **Final version of the Distance Randot Stereotest: normative data, reliability, and validity.** *J AAPOS.* 2010; **14**(2): 142–6.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Pearce KL, Sufrinko A, Lau BC, *et al.*: **Near Point of Convergence After a Sport-Related Concussion: Measurement Reliability and Relationship to Neurocognitive Impairment and Symptoms.** *Am J Sports Med.* 2015; **43**(12): 3055–61.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Maples WC, Hoenes R: **Near point of convergence norms measured in elementary school children.** *Optom Vis Sci.* 2007; **84**(3): 224–8.
PubMed Abstract | Publisher Full Text

14. Antona B, Barrio A, Barra F, *et al.*: **Repeatability and agreement in the measurement of horizontal fusional vergences.** *Ophthalmic Physiol Opt.* 2008; **28**(5): 475–91.
PubMed Abstract | Publisher Full Text

15. Brozek J, Simonson E, Bushard W, *et al.*: **Effects of practice and the consistency of repeated measurements of accommodation and vergence.** *Am J Ophthalmol.* 1948; **31**(2): 191–8.
PubMed Abstract | Publisher Full Text

16. Rouse MW, Borsting E, Deland PN, *et al.*: **Reliability of binocular vision measurements used in the classification of convergence insufficiency.** *Optom Vis Sci.* 2002; **79**(4): 254–64.
PubMed Abstract | Publisher Full Text

17. Long S, Leclerc S, Tinjust D, *et al.*: **Determining consistency and agreement of scores across two measurements of the visual system: Test-retest reliability.** *Med Sci Sports Exerc.* 2018; **50**(5S): 664.
Publisher Full Text

18. Giffard P, Daly L, Treleaven J: **Influence of neck torsion on near point convergence in subjects with idiopathic neck pain.** *Musculoskelet Sci Pract.* 2017; **32**: 51–6.
PubMed Abstract | Publisher Full Text

19. Moganeswari D, Thomas J, Srinivasan K, *et al.*: **Test Re-Test Reliability and Validity of Different Visual Acuity and Stereoacuity Charts Used in Preschool Children.** *J Clin Diagn Res.* 2015; **9**(11): NC01–5.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Ettinger U, Kumari V, Crawford TJ, *et al.*: **Reliability of smooth pursuit, fixation, and saccadic eye movements.** *Psychophysiology.* 2003; **40**(4): 620–8.
PubMed Abstract | Publisher Full Text

21. Klein C, Fischer B: **Instrumental and test-retest reliability of saccadic measures.** *Biol Psychol.* 2005; **68**(3): 201–213.
PubMed Abstract | Publisher Full Text

22. Broglio SP, Ferrara MS, Macciocchi SN, *et al.*: **Test-retest reliability of computerized concussion assessment programs.** *J Athl Train.* 2007; **42**(4): 509–14.
PubMed Abstract | Free Full Text

23. Pediatric Eye Disease Investigator Group: **Interobserver reliability of the prism and alternate cover test in children with esotropia.** *Arch Ophthalmol.* 2009; **127**(1): 59–65.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Stereo optical co: **Randot stereotest**. In: Stereo optical co., ed.; 1995.
Reference Source

25. Rowe F: **Clinical orthoptics**. 3rd ed. Chichester, West Sussex: Wiley-Blackwell; 2012.
Publisher Full Text

26. D'Agostino D: **Basic examination: Physiology of eye movements - measurement of ductions, versions, and vergences**. In: Scott W, D'Agostino D, Weingeist Lennarson L, editors. *Orthoptics and ocular examination techniques*. Baltimore: Williams & Wilkins; 1983.
Reference Source

27. Hurtt J, Rasicovici A, Windsor C: **Comprehensive review of orthoptics and ocular motility: Theory, therapy, and surgery**. 2nd ed. Saint Louis: The C.V. Mosby Company; 1977.
Reference Source

28. Bishop A: **Convergence and convergent fusional reserves - investigation and treatment**. In: Doshi S, Evans BJW, editors. *Binocular vision and orthoptics: Investigation and management*. Oxford: Butterworth-Heineman; 2001; 28–33.
Publisher Full Text

29. Scheiman M, Gwiazda J, Li T: **Non-surgical interventions for convergence insufficiency.** *Cochrane Database Syst Rev.* 2011; (3): CD006768.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Hayes GJ, Cohen BE, Rouse MW, *et al.*: **Normative values for the nearpoint of convergence of elementary schoolchildren.** *Optom Vis Sci.* 1998; **75**(7): 506–12.
PubMed Abstract | Publisher Full Text

31. Sutter P, Harvey L: **Vision rehabilitation: Multidisciplinary care of the patient following brain injury**. Boca Raton: Taylor & Francis Group; 2011.
Reference Source

32. Birch E, Williams C, Drover J, *et al.*: **Randot preschool stereoacuity test: Normative data and validity.** *J AAPOS.* 2008; **12**(1): 23–6.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Piano ME, Tidbury LP, O'Connor AR: **Normative Values for Near and Distance Clinical Tests of Stereoacuity.** *Strabismus.* 2016; **24**(4): 169–72.
PubMed Abstract | Publisher Full Text

34. Shrout PE, Fleiss JL: **Intraclass correlations: uses in assessing rater reliability.** *Psychol Bull.* 1979; **86**(2): 420–8.
PubMed Abstract | Publisher Full Text

35. Bland JM, Altman D: **Statistical methods for assessing agreement between two methods of clinical measurement.** *Lancet.* 1986; **1**(8476): 307–10.
PubMed Abstract | Publisher Full Text

36. Koo TK, Li MY: **A guideline of selecting and reporting intraclass correlation coefficients for reliability research.** *J Chiropr Med.* 2016; **15**(2): 155–63.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Johnston BC, Thorlund K, Schünemann HJ, *et al.*: **Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units.** *Health Qual Life Outcomes.* 2010; **8**(1): 116.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Saville DJ: **Multiple comparison procedures: The practical solution.** *Am Stat.* 1990; **44**(2): 174–80.
Publisher Full Text

39. R core team: **R: A language and environment for statistical computing**. Vienna, Austria: R foundation for statistical computing; 2015. 2015.
Reference Source

40. Shrier I: **Vision Tests in Concussion**. 2019.
http://www.doi.org/10.17605/OSF.IO/VB4W8

41. Moser RS, Schatz P, Grosner E, *et al.*: **One year test-retest reliability of neurocognitive baseline scores in 10- to 12-year olds.** *Appl Neuropsychol Child.* 2017; **6**(2): 166–71.
PubMed Abstract | Publisher Full Text

42. de Jongh E, Leach C, Tjon-Fo-Sang M, *et al.*: **Inter-examiner variability and agreement of the alternate prism cover test (APCT) measurements of strabismus performed by 4 examiners.** *Strabismus.* 2014; **22**(4): 158–66.
PubMed Abstract | Publisher Full Text

43. Mainwaring L, Pennock KMF, Mylabathula S, *et al.*: **Subconcussive head impacts in sport: A systematic review of the evidence.** *Int J Psychophysiol.* 2018; **132**(Pt A): 39–54.
PubMed Abstract | Publisher Full Text

44. Bujang MA, Baharum N: **A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review.** *Arch Orofac Sci.* 2017; **12**(1): 1–11.
Reference Source

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 5

Reviewer Report 10 September 2020

https://doi.org/10.5256/f1000research.29392.r71042

✓ **James P Dickey**
School of Kinesiology, University of Western Ontario, London, Ontario, Canada
Dillon Richards
Health and Rehabilitation Sciences, University of Western Ontario, London, Canada

The authors have revised the paper to acknowledge "Some data in these athletes appear to be outside the normative range of data previously described for the general population", and have provided access to the raw data through the data availability link. This enables the readers to evaluate the credibility fo the data and interpret the findings accordingly.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomechanics, head impact exposure in sports, concussion

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 4

Reviewer Report 02 September 2020

https://doi.org/10.5256/f1000research.28661.r70272

? **Dillon Richards**

Health and Rehabilitation Sciences, University of Western Ontario, London, Canada

James P Dickey

School of Kinesiology, University of Western Ontario, London, Ontario, Canada

Thank you for considering the points that we raised in the review, and we note that your recent revisions better characterize the effects of the outliers. We note that you have not chosen to acknowledge our point that these data points are extreme outliers (four data points at 3 or more IQRs above the third quartile, including one value 8.125 IQRs above the third quartile), and that they exceed the range of normative data for Positive Fusional Vergence that you present in Table 1. Our previous review stated that there is a strong reason for believing that these data points are questionable. In fact there is strong evidence that these data points should be eliminated rather than simply evaluating their influence in a "sensitivity analysis".

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomechanics, head impact exposure in sports, concussion.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 03 Sep 2020

Ian Shrier, McGill University, Montreal, Canada

Responses to reviewer 2&3:

Comment: Thank you for considering the points that we raised in the review, and we note that your recent revisions better characterize the effects of the outliers. We note that you have not chosen to acknowledge our point that these data points are extreme outliers (four data points at 3 or more IQRs above the third quartile, including one value 8.125 IQRs above the third quartile), and that they exceed the range of normative data for Positive Fusional Vergence that you present in Table 1. Our previous review stated that there is a strong reason for believing that these data points are questionable. In fact there is strong evidence that these data points should be eliminated rather than simply evaluating their influence in a “sensitivity analysis”.

Response: We thank the reviewers for their feedback. There are three points raised.

1. The reviewers insist that results excluding outliers be considered the primary analysis, and the results including all the data be considered secondary.
2. The reviewers suggest there are four outliers instead of the 2 outliers we noted.
3. The reviewers suggest that 4 points exceed the range of normative data we provided.

1. Sensitivity Analyses

We think that there are different ways to look at outliers and interpret the results. As we mentioned in our previous response to the reviewers, we think that “eliminating” data, as the reviewers suggest, is not the optimal approach. Instead, the sensitivity analysis, as we performed, is a preferable approach unless there is a clear data error. In this way, we are transparent about our data set and our analysis, and the readers can evaluate our findings.

F1000Research does not have an editor as an arbitrator when authors and reviewers disagree. Therefore, we have made the changes recommended by the reviewer and the analysis with the excluded data is now considered the primary result. We had already based our conclusions on the analyses after exclusions and have now edited the rest of the text as well.

2. Outliers

We are not sure why the reviewer thinks there are four outliers. The data in our study are available online at https://osf.io/gnjdm/. Here are the calculations for outliers, which we defined using the common standard: 1.5*IQR above the 3rd quantile.

Positive Fusional Vergence 30cm

First Test:
25%: 20
75%: 31.25
IQR: 11.25
1.5*IQR: 16.9
Outlier Threshold (75%+1.5*IQR): 48.2

Second Test
25%: 23.75
75%: 30
IQR: 6.25
1.5*IQR: 9.4
Outlier Threshold (75%+1.5*IQR): 39.4

There is only one person with values that should be considered as outliers for positive fusional vergence at 30cm (Id=14). This occurred for both tests (90 on the first test and 85 on the second test). The text now reads:

- "Given the very high ICC and the presence of an outlier that greatly increased the range of the values for the measure (known to increase ICC), we repeated the analysis excluding the outlier. This decreased the ICC from 0.93 to 0.53, and increased the LoA to ±43.5%. One of the reviewers for this paper has insisted that the analysis without the outlier be considered the primary analysis."

We also added a sentence to the figure legend:

- "When the analysis was repeated excluding the outlier to the far right, the ICC decreased to 0.53 and the LoA increased to 43.5%."

Positive Fusional Vergence 3m

First Test
25%: 17.5
75%: 25
IQR: 7.5
1.5*IQR: 11.3
Outlier Threshold (75%+1.5*IQR): 36.3

Second Test
25%: 17.5
75%: 21.25
IQR: 3.75
1.5*IQR: 5.6
Outlier Threshold (75%+1.5*IQR): 26.8

There are two people with values that should be considered as outliers for positive fusional vergence at 3m (Ids 14 and 15). Particiant 14 is an outlier for both measures, and Particicpant 15 is an outlier for the first test. When Participant 14 was removed the ICC dropped from 0.56 to 0.21 as we reported. If we remove only Participant 15, the ICC actually increases from 0.56 to 0.63. If we remove both outliers, the ICC was 0.45 and the LoA decreased from 60.2% to 41.4%. As per the reviewer's request, we are reporting the analysis with both outliers removed. The text now reads:

- "There were two outliers for Positive Fusional Vergence at 3m (one participant on both measures and one participant on only one measure). When we removed both of these outliers, the ICC dropped from 0.56 to 0.45 and the LoA decreased from 60.2% to 41.4%. In both of these cases, the two scores from the outlier were quite different. Although one might anticipate that the ICC would increase by removing such outliers, the ICC actually decreased because the range of values for the measure decreased substantially. As above, one of the reviewers for this paper insisted that the analysis without the outliers be considered the primary analysis."

We have also added a sentence to the figure legend that says:

- "When the analysis for Positive Fusional Vergence at 3m was repeated excluding the two outliers, the ICC decreased to 0.45 and the LoA decreased to 41.4%."

The results of these tests are also reported in the Discussion. The text in that section now reads:

- "When we removed the outlier for Positive Fusional Vergence at 30cm, the ICC dropped to 0.53, which is similar to the value found for the one-week test-retest reliability (ICC=0.54); the LoA increased to 43.5%. When we removed the two outliers from Positive Fusional Vergence at 3m, the ICC decreased to 0.45 and LoA decreased to 41.4%. Note that the outliers for this measure had large differences between the two test scores, and removing such data points would normally be expected to increase the ICC (Figure 3). The finding that the ICC decreased indicates that as expected, if the range of values among the populations is similar, the one-year test-retest reliability for Positive Fusional Vergence at both 30cm and 3m is likely less than the one-week test-retest reliability."

4. Normative Data Range

We do not know why some data were outside previously described normative data range. These are the data we received from the clinician doing the test as part of his regular clinical practice. It is possible that previously published normative data for the population does not represent normative data for athletes like those included in our study. We have added one sentence mentioning this in the limitations section of the article. It says:

- "Some of the data in these athletes appear to be outside the normative range of data previously described for the general population."

Competing Interests: No competing interests were disclosed.

Reviewer Report 27 August 2020

https://doi.org/10.5256/f1000research.28661.r70273

✓

M Nadir Haider
Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, USA

No additional comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Concussion, biomarkers, physiology, cerebral blood flow.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 3

Reviewer Report 20 July 2020

https://doi.org/10.5256/f1000research.27076.r66529

?

Dillon Richards
Health and Rehabilitation Sciences, University of Western Ontario, London, Canada

James P Dickey
School of Kinesiology, University of Western Ontario, London, Ontario, Canada

This is an interesting and important paper based on the prevalence of concussion and the growing appreciation vision tests for diagnosing and assessing concussion.

The outlying data points in Positive Fusional Vergence at 30 cm and 3 m have been identified by the previous external peer reviewers and warrant additional consideration. You present your ICC findings with and without the outlying data points, which is appropriate. However, you do not fully characterize the extreme deviance of the outlying data points. Your wording about the outliers is rather misleading – you state "we noticed one outlier that greatly increased the range of values along x-axis in Figure 2 and Figure 3". However, both the x- and y-coordinates of the outlier in Figure 2 meet your definition of outlier (1.5 interquartile ranges below the first quartile or above the third quartile), so it is not merely an issue with the x-axis. As well, you state "There was also one outlier for Positive Fusional Vergence at 3 m, 1.5 interquartile range above the third quartile", but the 1.5 interquartile range (IQR) is your threshold for identifying outliers, not the description of the outlier – this data point is actually 8.125 IQRs above the third quartile. Statistically speaking, it is extremely unlikely that this data point is part of the same distribution as the rest of the data set. Finally, and perhaps most importantly, you state that you "have no reason to believe the data are inaccurate". However, these outlying data points all exceed the range of normative data for Positive Fusional Vergence that you present in Table 1, providing a strong reason for believing that these data points are questionable. Your previous response "if deleting a point improved the ICC, we are confident the reviewer would agree that we should not delete the data point" trivializes the issue. The issues about the outlier data points must be more thoroughly addressed in the manuscript.

It is highly unfortunate that the sample size is so limited, particularly since it would appear that your inclusion criteria were quite broad (followed by the Institut National du Sport du Quebec from 2015–2018). Examination of the participants' durations between tests reveals that the majority of the participants had 335-336 or 371-371 days between assessments - presumably these dates correspond to the timing of the preseason tests for the different sports. Would you have more eligible participants if you had broadened the eligibility criterion?

It is unclear how it could be that your participants were limited to waterpolo and short-track speed skating, when presumably you started with a larger number of sports, but this should be clarified as it may reflect a bias in participant selection. As well, both waterpolo (Black *et al.* 2017)[1] and short-track speed skating (Quinn *et al.* 2003)[2] have a relatively high rate of concussions, and presumably the athletes may have received subconcussive head impacts, without receiving a concussion. Repetitive hits to the head are associated with microstructural and functional changes in the brain (Mainwaring *et al.* 2018)[3], and therefore should be acknowledged as a potential factor for the participants in this paper.

You identify that test-retest reliability of vision tests has been evaluated at the 1 day to 45 days time span. However, studies have evaluated longer-term test-retest reliability. For example, Klein and Fischer (2005)[4] evaluated 19-month test–retest correlations of pro- and anti-saccadic eye movements on 117 participants. Of more direct relevance to the student athletes evaluated in your paper, Breedlove *et al.* (2019)[5] evaluated the reliability of the King-Devick test (prosaccades) on NCAA athletes, including 833 participants with measures one year apart, and Naidu *et al.* (2018) [6] evaluated the season-to-season reliability of the King-Devick Test in Canadian professional football players. Your paper would be strengthened by incorporating a fuller complement of relevant papers that have performed longer-term test-retest reliability measures of vision tests, and comparing your findings with theirs.

The saccade measures reported in the paper have extremely limited value as they were collected using proprietary equipment - they should likely be removed from the paper.

The scatterplots (Figures 2A, 3A and 4A) show the line of identity, but it would be interesting to also see the line of best fit. Furthermore, for the parameters with outliers, it would be interesting to add the lines of best fit with and without the outlier.

The raw data presented through the Data Availability link is very helpful for gaining insight into the specifics of your data. However, it reveals that all of the data are reported as integers. Is this level of precision adequate for capturing the various vision tests? It would be helpful to include a "data dictionary", as recommended for best practices with spreadsheets (Broman and Woo, 2017)[7]
.

References

1. Black AM, Sergio LE, Macpherson AK: The Epidemiology of Concussions: Number and Nature of Concussions and Time to Recovery Among Female and Male Canadian Varsity Athletes 2008 to 2011.*Clin J Sport Med*. 2017; **27** (1): 52-56 PubMed Abstract | Publisher Full Text
2. Quinn A, Lun V, McCall J, Overend T: Injuries in short track speed skating.*Am J Sports Med*. **31** (4): 507-10 PubMed Abstract | Publisher Full Text
3. Mainwaring L, Ferdinand Pennock KM, Mylabathula S, Alavie BZ: Subconcussive head impacts in sport: A systematic review of the evidence.*Int J Psychophysiol*. **132** (Pt A): 39-54 PubMed Abstract | Publisher Full Text
4. Klein C, Fischer B: Instrumental and test-retest reliability of saccadic measures.*Biol Psychol*. 2005; **68** (3): 201-13 PubMed Abstract | Publisher Full Text
5. Breedlove KM, Ortega JD, Kaminski TW, Harmon KG, et al.: King-Devick Test Reliability in National Collegiate Athletic Association Athletes: A National Collegiate Athletic Association-Department of Defense Concussion Assessment, Research and Education Report.*J Athl Train*. 2019; **54** (12): 1241-1246 PubMed Abstract | Publisher Full Text
6. Naidu D, Borza C, Kobitowich T, Mrazik M: Sideline Concussion Assessment: The King-Devick Test in Canadian Professional Football. *Journal of Neurotrauma*. 2018; **35** (19): 2283-2286 Publisher Full Text
7. Broman K, Woo K: Data Organization in Spreadsheets. *The American Statistician*. 2018; **72** (1): 2-10 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomechanics, head impact exposure in sports, concussion.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 06 Aug 2020

Ian Shrier, McGill University, Montreal, Canada

Author Responses

REVIEWER # 2&3

Comment: This is an interesting and important paper based on the prevalence of concussion and the growing appreciation vision tests for diagnosing and assessing concussion.

Answer: We thank the reviewers for their interest in our manuscript.

Comment: The outlying data points in Positive Fusional Vergence at 30 cm and 3 m have been identified by the previous external peer reviewers and warrant additional consideration. You present your ICC findings with and without the outlying data points, which is appropriate. However, you do not fully characterize the extreme deviance of the outlying data points. Your wording about the outliers is rather misleading – you state “we noticed one outlier that greatly increased the range of values along x-axis in Figure 2 and Figure 3”. However, both the x- and y-coordinates of the outlier in Figure 2 meet your definition of outlier (1.5 interquartile ranges below the first quartile or above the third quartile), so it is not merely an issue with the x-axis. As well, you state “There was also one outlier for Positive Fusional Vergence at 3 m, 1.5 interquartile range above the third quartile”, but the 1.5 interquartile range (IQR) is your threshold for identifying outliers, not the description of the outlier – this data point is actually 8.125 IQRs above the third quartile. Statistically speaking, it is extremely unlikely that this data point is part of the same distribution as the rest of the data set. Finally, and perhaps most importantly, you state that you “have no reason to believe the data are inaccurate”. However, these outlying data points all exceed the range of normative data for Positive Fusional Vergence that you present in Table 1, providing a strong reason for believing that these data points are questionable. Your previous response “if deleting a point improved the ICC, we are confident the reviewer would agree that we should not delete the data point” trivializes the issue. The issues about the outlier data points must be more thoroughly addressed in the manuscript.

Answer: We thank the reviewers for raising this point and have modified the text accordingly. For the comment that 1.5 IQR is the threshold and not the data point, we agree and have removed the phrase. It now reads:

“There was also one outlier for Positive Fusional Vergence at 3m. When removing this outlier in a sensitivity analysis, the ICC dropped from 0.57 to 0.21.”

The paragraph in which this is mentioned refers to the fact that the 1-year test-retest reliability had higher ICC than the 1-week test-retest reliability and this should not be possible. Our sensitivity analyses were conducted to determine if this occurred because of the increased range observed in the 1-year data. The reviewers are correct that the outlier in question is indeed an outlier on both the x and y axis. We have modified the text accordingly. This particular section now reads as below. Similar changes were made to other parts of the manuscript where appropriate:

“Given the very high ICC and the presence of an outlier that greatly increased the range of values for the measure (known to increase ICC), we conducted a sensitivity analysis excluding the outlier.”

With respect to justifying keeping the outlier in the plot or not, we did not mean to trivialize the issue. We only meant that the decision to remove an outlier needs more justification than simply that the data point was unexpected. Therefore, although we agree with the reviewers that our sensitivity analysis for the ICC is more likely to be correct, we do not feel there is enough evidence to replace the original analysis with the sensitivity analysis as the primary analysis. We feel that discussing this at length would be more confusing than helpful and have deleted the phrase related to “have no reason to believe the data are inaccurate”. The full paragraph now reads:

“In one-year test-retest, Positive Fusional Vergence showed excellent reliability at 30cm (ICC=0.93) and moderate at 3m (ICC=0.56), initially. These values were better than the one-week test-retest reliability (ICC=0.54 and 0.49, respectively) [17] . It is difficult to understand how test-retest reliability over one year could be better than test-retest reliability over one week. When we explored the data further, we noticed one outlier that greatly increased the range of values for Positive Fusional Vergence at 30cm (Figure 2) and Positive Fusional Vergence at 3m (Figure 3). Increasing the range of values is known to increase the ICC. This is because ICC is based on the results of an analysis of variance which separates the error into variability between individuals (range of values along x or y axes) and variability within an individual. Therefore, if variability between persons increases, indicated by a larger range of values, ICC will increase. We explored how removing the outlier in our data would affect the results. When we removed the outlier for Positive Fusional Vergence at 30cm, the ICC dropped to 0.53, which is below the value found for the one-week test-retest reliability; it did not affect LoA. When we removed the outlier (same person) from Positive Fusional Vergence at 3m, the ICC decreased to 0.21. Note that the outlier for this measure had a large difference between the two test scores, and removing such a data point would normally be expected to increase the ICC (Figure 3). The finding that the ICC decreased indicates that as expected, if the range of values among the populations is similar, the one-year test-retest reliability for Positive Fusional Vergence at both 30cm and 3m is likely less than the one-week test-retest reliability.”

Comment: It is highly unfortunate that the sample size is so limited, particularly since it would appear that your inclusion criteria were quite broad (followed by the Institut National du Sport du Quebec from 2015–2018). Examination of the participants' durations between tests reveals that the majority of the participants had 335-336 or 371-371 days between assessments - presumably these dates correspond to the timing of the preseason tests for the different sports. Would you have more eligible participants if you had broadened the eligibility criterion?

Answer: We thank the reviewers for this comment. Our eligibility criteria only required that the athlete not have a concussion or undergo vision training between tests, and did not have a condition that would affect the results of vision testing. We are not sure which of these criteria the reviewers think we could relax and still obtain an unbiased answer to the question of 1-year test-retest reliability. We could have shortened the interval to only several months, but that would no longer be answering the 1-year test-retest reliability question. We have not made any changes to the manuscript.

Comment: It is unclear how it could be that your participants were limited to waterpolo and short-track speed skating, when presumably you started with a larger number of sports, but this should be clarified as it may reflect a bias in participant selection. As well, both waterpolo (Black et al. 2017) and short-track speed skating (Quinn et al. 2003) have a relatively high rate of concussions, and presumably the athletes may have received subconcussive head impacts, without receiving a concussion. Repetitive hits to the head are associated with microstructural and functional changes in the brain (Mainwaring et al. 2018), and therefore should be acknowledged as a potential factor for the participants in this paper.

Answer: We thank the reviewers for raising these points. For the types of sports participants were engaged in, these are the data provided to us. Many athletes from other sports only had 1 test, and some had concussions or vision testing within the 1-year interval. We do not have data on which athletes were referred for testing but never went for the test. The reviewers suggested Black et al reported waterpolo as a sport with many concussions. However, the study cited actually reported 0 concussions in waterpolo athletes. The Quinn et al study reported 6 concussions in 63 athletes over a 1-year period. The authors did not include the injury rate in the paper and it is not possible to compare risks to other sports without knowing how often the athletes were competing / practicing. In general, short-track speed skating concussions occur because of collisions that cause the athlete to fall, and then they may hit their head into the padded boards or on the ice. There are not multiple small hits like one would receive in American football or hockey. That said, we expand on the issue below for other studies that might include athletes from these types of sports.

The reviewers suggest cumulative subconcussive head impacts should be raised as potential factor for participants in this study. We are not sure what the reviewers mean. We agree with the paper by Mainwaring et al. (2018) that the reviewers cited. Mainwaring et al states:

- "Both the research and conceptual understanding of this phenomenon are in their infancy"
- "the findings are equivocal regarding the effect of subconcussive impacts on the brain"
- "Insufficient evidence was presented to conclude that repetitive head impacts are associated with neurocognitive impairment. It may be that neuropsychological assessment tools are not sufficiently sensitive to detect any subtle changes in cognitive function that emerge from subconcussive impacts, or that the neurocognitive changes are inconsequential, or follow neurophysiological changes or damage."
- "Future research is needed to characterize the phenomenon in question."

As an example, one study found that repetitive subconcussive head impacts over a single season do not appear to result in short-term neurologic impairment (see Gysland SM, Mihalik JP, Register-Mihalik JK, Trulock SC, Shields EW, Guskiewicz KM. *The relationship between subconcussive impacts and concussion history on clinical measures of neurologic function in collegiate football players*. Annals of biomedical engineering. 2012;40(1):14-22).

Aside from these results that do not support a decrease in neurocognitive function with subconcussive impacts, our objective in this study was to report on the 1-year test-retest reliability of vision tests in order to help clinicians understand how to interpret differences between testing conducted post-concussion and at baseline. If subconcussive impacts did affect vision testing, one would expect a decline in visual function as a consequence of subconcussive impacts. If this occurred, any change in test scores between baseline and post-concussion could not be attributed to concussion. That said, we doubt this is the case because if that were true, one would expect a decline in vision function (and test scores) conducted one year after baseline. We did not observe this in our data. However, we acknowledge that the athletes in our study were not involved in sports with many subconcussive impacts. We have modified the text at the end of the limitation, which now includes:

"This is a historical cohort observational study, a study design which has inherent limitations. The data provided were not always as precise as one might expect (e.g. near point convergence measured to the nearest cm). In addition, the sample size was relatively small and composed of healthy athletes, which will limit the generalizability of these findings to other populations. Although we started with a pool of 199 athletes, many athletes were excluded because they only had one baseline test, a concussion occurred in between the two baseline tests, or the second baseline test occurred outside the testing window of 365±30 days. Despite starting with athletes from many sports, only athletes from Waterpolo and Short-track speed skating met our eligibility criteria. It is unclear if subconcussion impacts affect neurological function in general[43]. If subconcussion impacts were common in these sports and affected vision testing, we should have seen a systematic decrease in vision capacity between the two tests; this was not observed. Further, if it were present, the effect would be considered part of the "noise" clinicians have to consider when comparing the results from post-concussion and baseline tests."

Comment: You identify that test-retest reliability of vision tests has been evaluated at the 1-45 days time span. However, studies have evaluated longer-term test-retest reliability. For example, Klein and Fischer (2005) evaluated 19-month test–retest correlations of pro- and anti-saccadic eye movements on 117 participants. Of more direct relevance to the student athletes evaluated in your paper, Breedlove et al. (2019) evaluated the reliability of the King-Devick test (prosaccades) on NCAA athletes, including 833 participants with measures one year apart, and Naidu et al. (2018) evaluated the season-to-season reliability of the King-Devick Test in Canadian professional football players. Your paper would be strengthened by incorporating a fuller complement of relevant papers that have performed longer-term test-retest reliability measures of vision tests, and comparing your findings with theirs.

Answer: We thank the reviewers for the reference that we had not been aware of. Our study investigated tests for specific visual function. Although the King-Devick test is sometimes used in concussion, it measures a combination of functions much beyond visual function. Therefore, we do not feel it is relevant to our research questions. We were not aware of the Klein and Fischer article and have now included the reference in the Introduction and Discussion. Our test was quite different from that studied in Klein and Fisher. The introduction text now reads:

Previous investigations of the test-retest reliability of these vision tests have used short test-retest time intervals ranging from 1 day to 45 days [9–17] , except for one test of saccades[44].

and the Discussion text now reads:

"However, we could not find any research examining the stability of the vision tests over a one year period, in athlete or non-athlete populations except for one test of saccades that was very different from the test used in this study[44]."

Comment: The saccade measures reported in the paper have extremely limited value as they were collected using proprietary equipment - they should likely be removed from the paper.

Answer: We respectfully disagree with the reviewers. First, we do not see any harm in

including the result of a non-standard test and readers who are not interested can simply ignore the results. Second, we evaluated this non-standard test as this measure was in our *a priori* protocol. Omitting analyses described in an *a priori* protocol is a form of reporting bias that we would prefer to avoid. We have modified the limitation section to say:
"Finally, the results of the test of Saccades in this study are based on the unpublished proprietary algorithm developed by the clinician. This limits its applicability for other clinicians."

Comment: The scatterplots (Figures 2A, 3A and 4A) show the line of identity, but it would be interesting to also see the line of best fit. Furthermore, for the parameters with outliers, it would be interesting to add the lines of best fit with and without the outlier.

Answer: We thank the reviewers for this comment. We believe the recommended statistical practice for evaluating reliability is the ICC with line of identity, and LOA. We have provided the references that guided this decision. Lines of best fit are not measures of reliability. In addition, any comparison of regression lines with the line of identity can be misleading because one must incorporate the uncertainty due to sampling. If the reviewers have an appropriate statistical reference that supports using regression in studies of test-retest reliability, we would be happy to add the analyses in a subsequent revision.

Comment: The raw data presented through the Data Availability link is very helpful for gaining insight into the specifics of your data. However, it reveals that all of the data are reported as integers. Is this level of precision adequate for capturing the various vision tests? It would be helpful to include a "data dictionary", as recommended for best practices with spreadsheets (Broman and Woo, 2017).

Answer: We thank the reviewers for their comment. We have developed a data dictionary and uploaded it as metadata. We agree that some of the measures could have been measured more precisely than others but these are the data provided from the clinician with expertise in orthoptics. We have added text to the beginning of the 2nd paragraph in the limitations section which now reads:

"This is a historical cohort observational study, a study design which has inherent limitations. The data provided were not always as precise as one might expect (e.g. near point convergence measured to the nearest cm)."

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 10 March 2020

https://doi.org/10.5256/f1000research.21476.r60656

✓ **M Nadir Haider**

Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, USA

Thank you for giving me the opportunity to review this manuscript. It measures the retest reliability of common ocular/oculomotor tests over one year. The sample size is 16 college-aged athletes. Intra-class correlation is performed and presented. I have read through the entire manuscript and it is exceptionally well-written, it shows that it has gone through several internal, and even some external, reviews and revisions already. The statistical analysis are correctly described and the appropriate tests and graphs are used to present data.

The most obvious downside of this study is the small sample size, there is so much within-subject variation among these test due to the natural process of aging and ocular adaptations which could be due to insignificant events like getting a new monitor for work. Future studies should be performed on larger sample sizes, etc.

But I believe that there is merit in having your study indexed for a couple of reasons. The research protocol and analysis are well explained and could be used for design future oculomotor retest reliability studies. Secondly, I am glad that you had concussion as your exclusionary criteria since there are a hundred different publications showing abnormalities in vision function tests after concussion, yet present no retest reliability without the presence of a concussive head injury. I think this paper provides some preliminary evidence which should be made available to other researchers and I think this is a citable manuscript. I do not have any sentence by sentence suggestions, but my only major suggestion is to remove the pre-outlier ICC of Positive Fusional Vergence at 30cm value of 0.93 and say that it is 0.55 (moderate). And I think Negative Fusional Vergence at 30cm should be classified as Good ICC (not moderate since it is between 0.75 and 0.9).

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical design, physiological and biochemical markers of concussion, autonomic regulation of cerebral blood blow.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 23 Mar 2020

Ian Shrier, McGill University, Montreal, Canada

REVIEWER #1

Comment: Thank you for giving me the opportunity to review this manuscript. It measures the retest reliability of common ocular/oculomotor tests over one year. The sample size is 16 college-aged athletes. Intra-class correlation is performed and presented. I have read through the entire manuscript and it is exceptionally well-written, it shows that it has gone through several internal, and even some external, reviews and revisions already. The statistical analysis are correctly described and the appropriate tests and graphs are used to present data.

Answer: We thank the reviewer for the kind comments.

Comment: The most obvious downside of this study is the small sample size, there is so much within-subject variation among these test due to the natural process of aging and ocular adaptations which could be due to insignificant events like getting a new monitor for work. Future studies should be performed on larger sample sizes, etc.

Answer: In this paper, we used all eligible participants from a clinical database. Therefore, we could not calculate an a priori sample size. Our primary approach to sample size requirements is to estimate precision rather than use hypothesis testing. We have tried to provide information for both approaches in the current version, and the new final paragraph of the limitations section is provided below.

"This is a historical cohort observational study, a study design which has inherent limitations. In addition, the sample size was relatively small and composed of healthy athletes, which will limit the generalizability of these findings to other populations. Although we started with a pool of 199 athletes, many athletes were excluded because they only had one baseline test, a concussion occurred in between the two baseline tests, or the second baseline test occurred outside the testing window of 365±30 days. With an effective sample size of 16, the anticipated precision of ICC estimates was +/- 0.25 and the study had 80% power to detect ICC values >= 0.6 and more than 90% power to detect ICC values >=0.7 i.e. rejection of the null hypothesis (Table 1a in [42]). Note that a total of >60 individuals were required to exclude ICC values <=0.5 with 80% power and an anticipated true ICC>0.7 (Table 2b in [42])."

Reference: Bujang MA, N. B. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. *Arch Orofac Sci*. 2017; 12(1): 1-11.

Comment: But I believe that there is merit in having your study indexed for a couple of reasons. The research protocol and analysis are well explained and could be used for design future oculomotor retest reliability studies. Secondly, I am glad that you had concussion as your exclusionary criteria since there are a hundred different publications showing abnormalities in vision function tests after concussion, yet present no retest reliability without the presence of a concussive head injury. I think this paper provides some preliminary evidence which should be made available to other researchers and I think this is a citable manuscript.

Answer: We again thank the reviewer for the kind comments.

Comment: I do not have any sentence by sentence suggestions, but my only major suggestion is to remove the pre-outlier ICC of Positive Fusional Vergence at 30cm value of 0.93 and say that it is 0.55 (moderate).

Answer: We thank the reviewer for the comment. Recommended practice is to only delete data points if you have a very good reason to believe they are inaccurate. Otherwise, one should keep the original analysis intact and apply sensitivity analyses. For example, if deleting a point improved the ICC, we are confident the reviewer would agree that we should not delete the data point. For this reason, we have not changed our results as suggested. However, we have modified the text to further emphasize the importance of the sensitivity analysis.

Comment: And I think Negative Fusional Vergence at 30cm should be classified as Good ICC (not moderate since it is between 0.75 and 0.9).

Answer: We thank the reviewer for pointing out this oversight. We have now indicated that Figure 3 shows results for good to moderate reliability tests, and made the associated changes in the abstract and manuscript as well.

Competing Interests: No competing interests were disclosed.