RESEARCH ARTICLE

One-year test-retest reliability of ten vision tests in Canadian athletes [version 1; peer review: 1 approved]

Mehdi Aloosh^{1,2}, Suzanne Leclerc³, Stephanie Long⁴, Guowei Zhong⁴, James M. Brophy⁵, Tibor Schuster⁴, Russell Steele⁶, Ian Shrier^{4,7}

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

²Department of Health Research Methods, Evidence, and Impact, Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Canada

³Institut National du Sport du Quebec, Montreal, Canada

⁴Department of Family Medicine, McGill University, Montreal, Canada

⁵Faculty of Medicine, McGill University, Montreal, Canada

⁶Department of Mathematics and Statistics, McGill University, Montreal, Canada

⁷Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, Canada

V1 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 increasingly being suggested for use 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. Therefore, 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, etc.). 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 0.55 when an outlier was excluded. There was moderate reliability in Negative Fusional Vergence at 30cm (ICC=0.78), Phoria at 30cm (ICC=0.68), Near Point of Convergence

Open Peer Review

Approval Status

	1	2
version 5		
(revision)
09 Sep 2020		view
version 4		
(revision)
26 Aug 2020	view	view
version 3		
(revision)
08 Jun 2020		view
version 2		
(revision)		
31 Mar 2020		
version 1		
09 Jul 2019
	view	

1. **M Nadir Haider** , State University of New York at Buffalo, Buffalo, USA
2. **Dillon Richards**, University of Western Ontario, London, Canada

break (ICC=0.65) and Saccade (ICC=0.56). The ICC for Positive Fusional Vergence at 3m (ICC=0.56) also decreased to 0.45 after removing one outlier. 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 $\pm 40\%$ to $\pm 90\%$.

Conclusions: Four tests had 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

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.

Copyright: © 2019 Aloosh M *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 1; peer review: 1 approved]** F1000Research 2019, 8:1032 <https://doi.org/10.12688/f1000research.19587.1>

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

Introduction

Concussion, a form of mild traumatic brain injury is a growing public health concern¹. Estimates suggest up to 3.8 million sport-related concussions occur annually in the United States, with 50% going unreported². United States emergency department visits for sports-related traumatic brain injuries have increased 60% over 2001–2009³. 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 tests⁶. Previous research has shown an association between concussion and eye movement¹. Concussion may therefore affect multiple aspects of vision, including saccades, pursuit, convergence, accommodation, and vestibulo-ocular reflex⁷. Some studies reported 50% to 90% incidence of visual symptoms, such as blurred vision and diplopia in individuals with concussion⁸. 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^{9–16}. 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 1 day to 45 days^{9–17}. 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 used¹⁸. 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 1 day to 45 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 ($\pm$ 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 3m¹⁰. Phoria was measured at 30cm and 3m using the prism and alternate cover test using the procedures described by the Pediatric Eye Disease Investigator Group¹⁹. To perform NPC and NPCb, we followed the Maples *et al.*, protocol¹³. We measured Gross Stereoscopic Acuity with the Randot Stereotest (Stereo Optical Co., Inc., Chicago, IL) according to the manufacturer's instructions²⁰. 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.

Analysis

We report the mean (SD) for continuous variables at baseline. We evaluated test-retest reliability using Intraclass Correlation Coefficient (ICC)²¹ and 95% limits of agreement (LoA)²². 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 reliability²³. 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, $[(T1 - T2) / \text{mean}(T1 \& T2)] * 100$ ^{22,24}. 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.

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.

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)³⁴. 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 3.4.3³⁵. 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)³⁶. Our analysis suggested one-year test-retest reliabilities ranging from poor to excellent among the ten vision tests. 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

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.

Figure 3. Vision tests with 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 Saccade. (B) Bland-Altman plot related to each test. See Figure 2 for explanation of abbreviations and scales.

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.

1 year. The range of measurements was between 14 and 45 diopters with one outlier at 90 diopters. LoA of the test was $\pm 41.9\%$. Given the very high ICC and the presence of an outlier that greatly increased the range of x-axis (known to increase ICC), we conducted a sensitivity analysis excluding the outlier. This decreased the ICC from 0.93 to 0.55, and increased the LoA to $\pm 43.5\%$.

Five tests showed 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=59.8%), and Saccades (ICC=0.57, LoA=24.3%). There was also one outlier for Positive Fusional Vergence at 3m, 1.5 interquartile range above the third quartile. When removing this outlier, the ICC dropped from 0.57 to 0.21. In this case, the two scores from the outlier were quite different. Although one might anticipate that the ICC would increase by removing such an outlier, the ICC actually decreased because the range of the x-values decreased substantially.

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.

Discussion

We found that the one-year test-retest reliability for 10 vision tests in young elite athletes ranged from excellent to poor. Positive Fusional Vergence at 30cm showed excellent reliability, while that of the others appeared to be moderate to poor. 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 0.82³⁷. 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. 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 year¹⁸.

Although there are no long-term reliability studies on vision tests, 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^{9–16}. For instance, one study reported an ICC of 0.65 for NPC in healthy individuals¹⁵, while other studies have reported excellent reliability for NPCb in healthy school children (ICC=0.94)¹⁶ and concussed athletes with normal vision (ICC = 0.92 to 0.97)¹². In addition, we recently examined one-week test-retest reliability of the same ten vision tests with the same methods as this current study in 20 young non-elite 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. 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 along x-axis in Figure 2 and Figure 3. Increasing the x-axis range 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 (x-range of values) and variability within an individual. Therefore, if variability between persons increases, indicated by a larger range of values on the x-axis, ICC will increase. We explored how removing the outlier in our data would affect the results even though we have no reason to believe the data are inaccurate. When we removed the outlier, the ICC dropped below the value found for the one-week test-retest reliability but it did not affect LoA. A similar effect was observed when we removed an outlier from Positive Fusional Vergence at 3m, even though this particular outlier had a large difference between the two test scores, which would normally be considered to decrease ICC (Figure 3). This finding indicates that as expected, if the range of values among the populations is similar, the one-year test-retest reliability for Positive Fusional Vergence is likely less than the one-week test-retest reliability.

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 increased ICC could have occurred simply by chance because of sampling variation.

Negative Fusional Vergence tests showed good and poor one-year test-retest reliability, at 30cm and 3m (ICC=0.78 and 0.0, respectively). These results were similar to our one-week test-retest study¹⁷. Our results were also similar to those of another study examining one-week test-retest reliability of Positive Fusional Vergence reporting ICCs of 0.53–0.59¹⁶, which is similar to the ICC we found at 3m. The different results might occur because of minor differences in application of tests which are not explicitly mentioned in the *Methods*.

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 strabismus³⁸ or esotropia¹⁹ 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 measured³⁸. 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.* reported a similar ICC of 0.65 for NPC in healthy adults^{15,16}. However, other studies have reported excellent reliability for NPCb in school children (ICC=0.94)¹⁶ and in concussed athletes with normal vision (ICC=0.92 to 0.97)¹². The discrepancies in results are most likely due to differences in testing procedures. For instance, we used the Maples method¹³ which is a non-accommodative test, whereas other studies used an accommodative target, such as RAF rule²⁸, or Astron International Accommodative Rule^{12,16}. Giffard *et al.* found that RAF rule had good test-retest reliability (ICC=0.84) for NPC using within a one-week interval. The Astron International Accommodative Rule for NPCb had excellent one-week test-retest reliability (ICC=0.94–0.98) among healthy children^{16,39}.

Our one-year test-retest results for Gross Stereoscopic Acuity in young athletes showed poor reliability (ICC=0.03; 95% LoA= $\pm 92.5\%$) even though our previous one-week test-retest results reported good reliability in non-athlete young adults (ICC=0.86; 95% LoA = $\pm 55\%$)¹⁷ and another study using Titmus stereo fly and Frisby stereo tests in children revealed an excellent reliability (ICC=1.0)⁴⁰. In addition, one 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 children¹¹. With an 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^{11,17,40}.

Finally, our clinician's test of Saccades showed moderate reliability (ICC=0.57) with the smallest LoA (in percentage) of other tests, similar to the one-week study¹⁷. These results are similar to other findings in healthy adults over a two-month period (ICC=0.59)⁴¹. 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, Saccade 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^{19,38}. 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.

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.

Conclusion

We found that four out of the ten vision tests (Negative Fusional Vergence at 30cm, Phoria at 30cm, NPCb and Saccade) had moderate one-year test-retest reliability. This level of reliability is useful in conditions which produce a moderate change in vision function. The remaining six 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>³⁶

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

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.

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 (Apex Inc. and Varitron Technologies Inc.)

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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 Apex Inc. for examining the athletes.

References

- 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](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- Ventura RE, Balcer LJ, Galetta SL: **The neuro-ophthalmology of head trauma.** *Lancet Neurol.* 2014; **13**(10): 1006–16. [PubMed Abstract](#) | [Publisher Full Text](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- Pearce KL, Sufirinko 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](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- Stereo optical co: **Randot stereotest.** In: Stereo optical co., ed.; 1995.
- Shrout PE, Fleiss JL: **Intraclass correlations: uses in assessing rater reliability.** *Psychol Bull.* 1979; **86**(2): 420–8. [PubMed Abstract](#) | [Publisher Full Text](#)
- 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](#)
- 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](#)
- 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](#)
25. Rowe F: **Clinical orthoptics**. 3rd ed. Chichester, West Sussex: Wiley-Blackwell; 2012.
[Reference Source](#)
 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.
 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.
 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-Heinemann; 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](#)
 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. Saville DJ: **Multiple comparison procedures: The practical solution**. *Am Stat*. 1990; **44**(2): 174–80.
[Publisher Full Text](#)
 35. R core team: **R: A language and environment for statistical computing [Internet]**. Vienna, Austria: R foundation for statistical computing; 2015. 2015.
 36. Shrier I: **Vision Tests in Concussion**. 2019.
<http://www.doi.org/10.17605/OSF.IO/VB4W8>
 37. 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](#)
 38. 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](#)
 39. 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](#)
 40. 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](#)
 41. 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](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 10 March 2020

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

© 2020 Haider M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 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.

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.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research