Sport-Related Concussion and Vision Tests

Mehdi Aloosh, MD, MSc

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

Thesis submission: August 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master's in Epidemiology

Table of contents

Abstract

Background: Vision tests are noninvasive tests increasingly suggested for concussion assessment. To know if a test result is abnormal after a concussion, one needs to compare the result to a value obtained before a concussion (baseline). The test results may be affected by physiological and psychological changes of the patient over time and variations in how the clinician conducts the examination. Therefore, one needs to know how much the results may vary even without a concussion in order to determine if the change with a concussion (signal) is more than the expected change due to normal variability (noise). However, only a limited number of studies have assessed the test-retest reliability of vision tests, and none have evaluated test-retest time intervals longer than about 57 days. Because concussions may occur months after a yearly baseline test, appropriate interpretation of results can only be made if we understand one-year test-retest reliability.

Objective: Our objective was to determine the one-year test-retest reliability of ten vision tests in a cohort of healthy Canadian elite athletes, who did not suffer a concussion during one-year. These ten tests measured different aspects of visual function, including Positive Fusional Vergence at 30cm and 3m, Negative Fusional Vergence at 30cm and 3m, Phoria at 30cm and 3m, Near Point of Convergence and Near Point of Convergence break, Gross Stereoscopic Acuity, and Saccades.

Methods: We studied elite Canadian athletes followed at the Institut National du Sport du Quebec (INSQ), evaluated by a single INSQ sports medicine physician who was responsible for all vision testing referrals. The vision test data was abstracted from the medical records of a

single clinician trained in orthoptic testing (APEXK) performing the vision tests. The two data sets were linked, cleaned and harmonized according to a pre-determined set of rules. After data verification, we included athletes who completed two baseline evaluations within 365±30 days. We excluded athletes with any concussion, vision training in between the annual evaluations, or medical conditions/treatment that might affect the tests. We evaluated test-retest reliability using Intraclass Correlation Coefficient (ICC) and 95% limits of agreement (95% LoA). 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.

Results: There were 16 athletes, nine females and seven males, out of 199 who met our inclusion criteria, 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, 95% LoA=±41.9%), but the ICC dropped to 0.53 (95% LoA= ±43.5%) when an outlier was excluded in a sensitivity analysis. There was good to moderate reliability in Negative Fusional Vergence at 30cm (ICC=0.78, 95% LoA=±41.2%), Phoria at 30cm (ICC=0.68, 95% LoA=±119.2%), Near Point of Convergence break (ICC=0.65, 95% LoA=±49.4%) and Saccades (ICC=0.61, 95% LoA=±24.3%). The ICC for Positive Fusional Vergence at 3m (ICC=0.56, 95% LoA=±60.2%) also decreased to 0.21 after removing one outlier. We found poor reliability in Near Point of Convergence (ICC=0.47, 95% LoA=±73.9%), Gross Stereoscopic Acuity (ICC=0.03, 95% LoA=±92.5%) and Negative Fusional Vergence at 3m (ICC=0.0, 95% LoA=±48.4%). ICC for Phoria at 3m was not appropriate because scores were identical in 14/16 athletes.

Conclusions: Five vision tests had good to moderate one-year test-retest reliability after removing outliers, and the remaining tests had poor reliability in this healthy athlete population. These results represent the “noise” of the tests. Future research is needed to investigate if

concussions produce a signal that is greater than the noise. If this is the case, these vision tests will be useful clinically regardless of the absolute ICCs and LoA values.

Resume

Contexte: Les tests de vision sont des tests non invasifs de plus en plus suggérés pour l'évaluation des commotions cérébrales. Pour savoir si un résultat de test est anormal après une commotion cérébrale, il faut comparer le résultat à une valeur obtenue avant une commotion cérébrale (référence). Les résultats du test peuvent être affectés par des changements physiologiques et psychologiques du patient au fil du temps et par des variations dans la façon dont le clinicien conduit l'examen. Par conséquent, il faut savoir dans quelle mesure les résultats peuvent varier même sans commotion cérébrale afin de déterminer si le changement avec une commotion cérébrale (signal) est plus que le changement attendu en raison de la variabilité normale (bruit). Cependant, seulement un nombre limité d'études ont évalué la fiabilité de test-retest des tests de vision et aucune n'a évalué les intervalles de temps de test-retest supérieurs à une semaine. Étant donné que les commotions cérébrales peuvent survenir des mois après un test de référence annuel, une interprétation appropriée des résultats ne peut être faite que si nous comprenons la fiabilité de test-retest sur un an.

Objectif: Notre objectif était de déterminer la fiabilité de test-retest sur un an de dix tests de vision dans une cohorte d'athlètes canadiens d'élite en bonne santé, qui n'ont pas subi de commotion cérébrale au cours de la période d'un an. Ces dix tests ont mesuré différents aspects de la fonction visuelle. Ils ont inclus la Vergence fusionnelle positive à 30 cm et à 3 m, la Vergence fusionnelle négative à 30 cm et à 3 m, Phoria à 30 cm et à 3 m, le point de convergence proche et le point de rupture de convergence, l'acuité stéréoscopique grossière, et les Saccades.

Méthodes: Nous avons étudié les athlètes canadiens d'élite suivis à l'Institut National du Sport du Québec (INSQ), évalués par un seul médecin du sport de l'INSQ qui était responsable de toutes les références aux tests de vision. Les données du test de vision ont été extraites des dossiers médicaux d'un seul clinicien formé aux tests orthoptiques (APEXK) effectuant les tests de vision. Les deux ensembles de données ont été liés, nettoyés et harmonisés selon un ensemble de règles prédéterminées. Après vérification des données, nous avons inclus les athlètes qui ont effectué deux évaluations de base dans un délai de 365±30 jours. Nous avons exclu les athlètes ayant subi une commotion cérébrale, un entraînement visuel entre les évaluations annuelles ou des conditions/traitements médicales susceptibles d'affecter les tests. Nous avons évalué la fiabilité de test-retest en utilisant le coefficient de corrélation intraclasse (CCI) et des Limites de l'Accord à 95% (95% LdA). Nous avons considéré un CCI ≤0.5 comme faible, 0.51-0.74 comme modéré, 0.75-0.89 comme bon et ≥0.90 comme une excellente fiabilité.

Résultats: Il y avait 16 athlètes, neuf femmes et sept hommes, sur 199 qui répondaient à nos critères d'inclusion, avec un âge moyen de 22.7 (ET 4.5) ans. Parmi les tests de vision, nous avons observé une excellente fiabilité de test-retest en Vergence fusionnelle positive à 30 cm (CCI=0.93, 95% LdA=±41.9%), mais l'CCI a chuté à 0.53 (95% LdA=±43.5%) en cas de valeur aberrante a été exclu d'une analyse de sensibilité. Il y avait une fiabilité bonne à modérée dans la Vergence fusionnelle négative à 30cm (ICC=0.78, 95% LdA=±41.2%), Phoria à 30cm (CCI=0.68, 95% LdA=±119.2%), le point de rupture de convergence (CCI=0.65, 95% LdA=±49.4%) et Saccades (CCI=0.61, 95% LdA=±24.3%). L'CCI pour la Vergence fusionnelle positive à 3m (CCI=0.56, 95% LdA=±60.2%) a également diminué à 0.21 après élimination d'une valeur aberrante. Nous avons trouvé un manque de fiabilité au point de convergence proche (CCI=0.47, 95% LdA=±73.9%), à l'acuité stéréoscopique brute (CCI=0.03, 95%

LdA=±92.5%) et à la vergence fusionnelle négative à 3m (CCI=0.0, 95% LdA=±48.4%). CCI pour Phoria à 3 m n'était pas approprié car les scores étaient identiques chez 14/16 athlètes.

Conclusions: Cinq tests de vision avaient une fiabilité de test-retest de bonne à modérée sur un an après élimination des valeurs aberrantes et les tests restants avaient un manque de fiabilité dans cette population d'athlètes en bonne santé. Ces résultats représentent le bruit des tests. Des recherches futures sont nécessaires pour déterminer si les commotions cérébrales produisent un signal supérieur au bruit. Si tel est le cas, ces tests de vision seront cliniquement utiles quels que soient les CCI absolus et les valeurs LdA.

List of Tables

List of Figures

Abbreviations

TBI: Traumatic Brain Injury

SRC: Sport-Related Concussion

NACRS: National Ambulatory Care Reporting System

BVT: Binocular Vision Test

PCS: Post-Concussion Syndrom

GCS: Glasgow Coma Scale

INSQ: Institut National du Sport du Quebec

ICC: Intraclass Correlation Coefficient

LoA: Limits of Agreement

PEDIG: Pediatric Eye Disease Investigator Group

Acknowledgements

First and foremost, I greatly appreciate my thesis supervisors, Professor James Brophy and Dr. Ian Shrier, for their sincere support to succeed in this study when I was simultaneously doing a residency in public health and preventive medicine, including family medicine. I felt privileged to work with them and enjoyed every moment of my interaction with them. I would like to appreciate Dr. Ninh Tran and Dr. Julie Emili, public health and preventive medicine residency program directors at McMaster University, to support my master's study at McGill University. In addition, I would like to thank Dr. Vicki Tagalakis for the examination of my thesis. I would also like to thank Dr. Elaheh Ghodsi for her sincere assistance with the abstract's French translation. Finally, I would like to thank Dr. Arash Aloosh for the enlightening discussions.

On a personal note, I would like to express my deep gratitude to my parents, who taught me to be diligent, patient and humble in the path to success. There are no words to express my love for Hoda. She has inspired me to follow my passion since we know each other. I have been delighted by our son's birth when I was working on this thesis. He accompanied me in writing parts of my thesis. And he brought me additional inspiration to protect the next generations' health and well-being.

Conflict of Interest

I have no conflict of interests and receive no funding to perform this research.

Preface and Contribution

The original research idea was conceived by Suzanne Leclerc from the Institut National du Sport du Quebec, and David Tinjust from APEXK Inc. Ian Shrier was the methodologist for the study and guided the protocol and data analysis. Mehdi Aloosh contributed to tidying the data for his particular project. All non-manuscript chapters of this thesis were written by Mehdi Aloosh and critically reviewed and revised by James Brophy and Ian Shrier. The manuscript was written by Mehdi Aloosh, Suzanne Leclerc, Stephanie Long, Guowei Zhong, James Brophy, Tibor Schuster, Russell Steele, and Ian Shrier.

1 Introduction

1.1 Thesis rational

A concussion is a common type of injury and growing public health concern (1, 2). The incidence rate of concussion in Canadian children younger than 14 years old is 200 per 100,000 (3). In 2016-17, the National Ambulatory Care Reporting System (NACRS) of Canada identified approximately 46,000 concussions diagnosed in emergency departments among children and youth 5-19 years of age. Of these, 57% were male. In males, 54% were sport-related and in females, 45% were sport-related. The other causes were assaults, self-harm, and other unintentional non-sport-related reasons (4). Between 2005 and 2009, children had more than 2 million outpatient visits and almost 3 million emergency department visits concussion in the United States (5). In addition, it has been reported as many as about 50% of players did not report their concussion because they did not think the injury was serious (66%), they were motivated to continue competition (41%), or they were not aware of probable concussion (36%) (6). It also has been reported that some athletes do not want to be removed from play, despite formal education about the risks of concussion (7).

Concussion has been associated with long-term sequelae and a decrease in quality of life (8, 9). A concussion can impair balance, cognitive, vestibular, and oculomotor function (10, 11). It can also cause headaches, dizziness, visual disturbances, and other symptoms that can negatively affect performance in sport, school, and work (8, 9, 12). Quantifying the burden of concussion in Canada is difficult because data collection and reporting is not consistent across the country (13). In British Columbia in 2010 alone, the direct healthcare costs for concussion-

related hospitalizations alone were $2.4 million (14). Currently, the diagnosis of concussion and decisions to return-to-play are based on signs, symptoms, physical examination and special tests (10). A concussion may affect multiple aspects of vision, such as saccades (rapid eye movements that abruptly alter the point of fixation), pursuit (eye movement to closely follow a moving object), convergence (simultaneous inward movement of both eyes to maintain binocular vision), accommodation (eye reflex acting to maintain focus on near objects), and vestibulo-ocular reflex (a reflex acting to stabilize gaze during head movement) (15). Therefore, many studies have proposed that visual performance measures might enhance the diagnosis of concussion (15-17).

Binocular vision tests and Saccades are noninvasive tests with rapid administration and scoring of vision motor function. Each test measures a vision function that is linked to a particular brain structure or pathway (15). To know if a test result is abnormal after a concussion, one needs to compare the result to a value obtained before a concussion (baseline). The test results may be affected by physiological and psychological changes of the patient over time and variations in how the clinician conducts the examination. Therefore, one needs to know how much the results may vary even without a concussion in order to determine if the change with a concussion (signal) is more than the expected change due to normal variability (noise).

However, only a limited number of studies have determined test-retest reliability in binocular vision tests and Saccades (18-29). In addition, these studies are not uniform in their method and differ in the populations they studied. These studies also only assessed short term reliability, with test-retest time intervals ranging from 0 to 57 days (18-29). These results provide information on the usefulness of the vision tests when following improvement or deterioration of concussed patients for short periods. However, baseline tests are usually conducted once a year (pre-season) or less frequently (10, 30, 31). This means the concussion can occur several months

and up to one year after annual baseline testing. For test-retest reliability to be useful in determining vision motor dysfunction of a concussion, the time intervals must reflect the time frame in which these tests would potentially be used for comparison of a post-concussion test to a baseline test (32).

1.2 Thesis objective

The thesis objective is to evaluate the test-retest reliability of ten vision tests in a well-defined population of young Canadian athletes over one year. This study population was chosen as it is representative of a likely target population as most concussions occur in young people and are often sport related. The ten tests each measured different aspects of visual function. They included Positive Fusional Vergence at 30cm and 3m, Negative Fusional Vergence at 30cm and 3m, Phoria at 30cm and 3m, Near Point of Convergence, Near Point of Convergence break, Gross Stereoscopic Acuity, and Saccades.

2 Review of Literature

2.1 Traumatic Brain Injury

Traumatic brain injury (TBI) is a common medical problem. Almost 50% of Americans have had at least one TBI in their lifetime (33). TBI is also a significant cause of morbidity and mortality worldwide (33). Falls are the primary cause of TBI, especially in children and the elderly (34). TBI can be classified based on the location of the injury, causal mechanism and severity. Severity is usually classified as mild, moderate and severe (15). Clinicians use the Glasgow Coma Scale (GCS), a 15-point scale assessing verbal, motor and eye reactions to stimuli to assess the severity of TBI (35). A lower score suggests a more severe TBI. Mild TBI is limited to scores of 13 or higher (36).

Mild TBIs composed up to 90% of TBIs (37). Loss of consciousness is not common in mild TBI. When loss of consciousness occurs, it is less than 30 min (38). Mild TBI is very common among children. One study suggested that one in five children will sustain a mild TBI before the age of 16 years (39). In Canada and the US, management of mild TBI occurs mainly outside of hospitals. However, the emergency department is often the entry point into the health care system to manage the injury (40-42).

2.2 Concussion

A concussion is the most common form of mild TBI (10). Concussions and mild TBI have

been used in the literature interchangeably (43). Concussion or uncomplicated mild TBI, a subset of TBI, has no intracranial abnormalities related to trauma that could be found on conventional structural neuroimaging (10). A concussion reflects a functional disturbance rather than a structural injury. It causes acute transient neurological dysfunction (10, 44, 45). The most frequent symptoms of concussion include headaches (25-58%), cognitive issues (26-44%), dizziness (18-27%), neck pain (28%), difficulty sleeping (24-26%), fatigue (40%) and visual disturbances (16-21%) (12). Signs and symptoms due to a concussion typically resolve in 1 to 2 weeks in adults (46, 47). However, in youth athletes, signs and symptoms of Sport-Related Concussion (SRCs) may last longer. For instance, in one study, at least 25% of high school football players required up to 4 weeks to achieve recovery (48).

Widely accepted first-line management of an acute concussion is relative rest for the first 24-48 hours after a concussion (10). Relative rest regards not participating in any sports or any other forms of exercise (49). This is to reduce the metabolic demand of the brain and to reduce symptoms (10, 50, 51). After 24-48 hours of relative rest, patients should gradually return to their normal daily physical and cognitive activities without generating symptoms (10, 50). Clinicians assess clinical recovery by evaluating if symptoms are resolving, normalization of physical examination, and tolerance for one's normal daily activities (10). Education of the patient and/or caregivers on concussion and the management plan is an integral part of concussion management (2, 51, 52). Other recommended interventions to manage symptoms associated with a concussion include treatment for cervical or vestibular dysfunction and CBT for mood or behaviour issues (10).

Post-Concussion Syndrome (PCS) is the presence of longstanding symptoms after concussion beyond the generally accepted time for recovery (53). There is no consensus on how to define "longstanding" in PCS (53). The arbitrary threshold is generally considered more than four weeks in athletes (54). The DSM-IV has six diagnostic criteria for PCS, which includes persisting symptoms for more than three months (55). It is not clear why some people experience a longer recovery time. One systematic review has identified the following risk factors associated with prolonged recovery, which included studies on PCS: severity of acute and subacute symptoms following a concussion, subacute problems with headaches or depression, preinjury history of mental health issues, being a teenager and being female (56).

Several authors have reported other important consequences of SRC. Some authors believe children and adolescents may experience diffuse cerebral swelling, a potentially fatal condition if they suffer subsequent head trauma while still symptomatic (57). Other authors believe that intense exercise in a brain with vulnerability from a previous concussion can cause brain swelling and collapse without any subsequent head trauma (58). In addition, experiencing post-concussion syndrome earlier in life might affect mental and physical health significantly later on (12). Some studies suggest that having a concussion during athletic activities always increases the risk of subsequent concussions (59). However, these studies suffered from obvious confounding factors. In the only study to adjust for this confounding, the first concussion did not increase the risk of a subsequent concussion in patients who were managed appropriately (60).

2.3 SRC Diagnosis

Clinical signs and symptoms of SRC may rapidly change in the acute phase, which reflects the underlying physiological brain injury. Some signs and symptoms may present only after a period of time. Therefore, serial assessments are often necessary regardless of a primary normal evaluation (10). Currently, there is no test or marker to diagnose a concussion. A definitive diagnosis of SRC and decisions regarding when to safely return athletes to play is still a challenge to clinicians (10). The diagnosis of acute SRC involves a detailed history and assessment of a range of signs and symptoms, which are not a consequence of drug, alcohol, or medication, other injuries, such as cervical injuries, or other comorbidities, such as psychological factors (10). These signs and symptoms include (10):

1. Symptoms: Somatic, such as headache: cognitive such as being in a fog; and emotional symptoms, such as mood liability
2. Physical signs, such as loss of consciousness, amnesia, neurological deficit
3. Balance impairment, such as unsteady gait
4. Behavioural changes, such as irritability
5. Cognitive impairment, such as slowed reaction times
6. Sleep disturbance

As a basic rule, athletes with clear signs of SRC, or suspected to have SRC should be removed from play (61). The first line evaluation of a suspected SRC is an assessment by a health professional (62). Clinicians use an abbreviated clinical examination and sideline tests to immediately determine whether an athlete with a suspected concussion requires a more comprehensive evaluation. This comprehensive evaluation should be performed in a detraction-

free environment (10, 54). The Sport Concussion Assessment Tool (SCAT; its most recent version is SCAT5) is the current standard for the acute evaluation of suspected concussion (63). It includes a series of tests including memory assessment Maddocks questions, Glasgow Coma Scale (GCS), cervical spine assessment, cognitive examination, neurologic examination, and delayed recall (63)

Concussion-induced visual dysfunction has only been recognized relatively recently and is not formally part of the SCAT5 (64). A concussion is associated with many neuro-ophthalmologic signs and symptoms (17). Approximately 50% of the brain circuitry is related to vision, and eye movement requires a sophisticated interplay between both cortical and subcortical structures in the brain (1). Eye movements are involved in multiple visual functions, including maintaining fixation and tracking moving objects. The cognitive control of vision in eye movements needs the coordination of reflexive and voluntary activity via pathways that are vulnerable to injury in all forms of TBI (15). Ocular symptoms and/or dysfunction following SRC are common, occurring in 30–60% of patients with a concussion (65). One study found 69% of 100 subjects 11–17 years old with concussion had at least one visual problem, such as problems in saccades, pursuit, convergence, or accommodation (66). Recently, many studies have proposed that visual performance measures could enhance the diagnosis and management of concussion (15-17). Before reviewing some of the existing vision tests that could be used to assess a concussion, I will discuss two important concepts in clinical testing, “reliability” and “Limits of Agreement”.

2.4 Test-Retest Reliability and Agreement

The reproducibility of a test result is assessed through both reliability and agreement (67). Reliability describes how a measurement method or tool performs in detecting real variability between measurements (68) and includes information about the ability of test scores to distinguish between participants (69). The agreement provides information on the absolute degree of measurement error (68). In the context of concussion, both reliability and agreement are necessary to understand the utility of the tests. To fully understand these concepts, I review different sources of variability in the measurement in the following.

A test result consists of true measurement ± error (70), where the error may be systematic or random (70). Systematic errors are predictable and constant, and they create bias (70). For example, a "learning effect" means that performance on a second test is expected to be better than performance on an initial test even when there is no true improvement in the underlying construct the test is supposed to be measuring. In the context of vision testing, a patient who remembers the order of the letters associated with an eye chart will do better on a subsequent test even if their vision had not improved.

Random measurement errors are not predictable or constant. They may be due to 1) the measurement tool, 2) participant variability (e.g. random physiologic or mental changes that vary affect a participant's result at the time of testing), or 3) rater variability (e.g. random changes in the examiner's judgment or the way the examiner performs the tests). In truly random error, the error will have a zero average by definition. In addition, the error terms are independent (71).

This implies that the mean score of the infinite test scores of an individual is equal to the true score (no bias). Random error represents the "noise" in the test result. If there is more noise, it is more difficult to determine if a change in the test scores (signal) represents a true change in condition, i.e. a lower signal-to-noise ratio makes the test less useful. For example, in telephone communication, interference in the telephone line can create static (noise). For the voice (signal) message to be understood, the voice must be higher than the static. Whenever the signal to noise ratio becomes too low, it will be difficult to hear the message. In the context of vision testing and concussion, poor reliability/agreement means we would need greater changes in the underlying brain function before we could be confident that a concussion affected vision or that vision improved over time after a concussion.

In this thesis, I focus on reliability and agreement parameters for continuous variables because all tests were measured by continuous variables. Other forms of reliability and agreement for nominal and categorical parameters (e.g. Cohen's kappa) are beyond the scope of this thesis. For continuous variables, the suitable measures of reliability and agreement are Intraclass Correlation Coefficient (ICC) and Limits of Agreement (LoA), respectively. ICC determines to what degree repeated measurements are correlated and maintain their rank in a sample over repeated measurements (70). LoA informs about the variations in repeated measurements (72). These are further explained in the next section. Although a comparison of the means of two continuous measures (e.g. using the t-tests) is sometimes used, it provides information about differences between the means of the two sets of variables, but not individuals. Finally, the correlation coefficient is also sometimes used, but it only provides information about the association of the variables (strength of linear relationship) (73). For example, if the second

test is always twice the value of the first test, the correlation coeffficient will be 1 even though the test-retest reliability is poor and there is little agreement.

2.4.1 Intraclass Correlation Coefficient

The ICC measures to what extent two measurements are correlated (range 0 to 1) with the linearity of the relationship between two repeated measures. This represents how well the rank order of participants in one measurement is replicated in a second measurement. (74) Although there are different types of the ICC, it generally represents a ratio of the (a) between measurement variance, which is the change in the observed result that one is interested in (signal), and (b) the total variance in the observed result, which includes error (signal plus noise). If there is no error, then the numerator and denominator are equal, and ICC = 1. If the variance due to error increases, the denominator increases, which means the ICC decreases. Values closer to one represent the higher reliability, and one represents perfect reliability. For example, when ICC is 0.8, 80% of the total variance in vision test scores occurs because of true score changes, with the remaining 20% occurring because of error.

Each type of ICC requires different calculations for the variance of the error. Therefore, different ICCs may provide different results when applied to the same data given distinct assumptions. In fact, each type of ICC is appropriate for specific situations defined by study design (72). The most commonly used six types of ICC are described by Shrout and Fleiss (75), although other authors have proposed additional types of ICC using a similar framework (72). The six types of ICC consist of three different models, with each model having two different forms. The three different models are:

Model 1: Raters are randomly selected from a larger population of raters. Each participant is rated by only one rater.

Model 2: Raters are randomly selected from a larger population of raters. All participants are measured by all raters.

Model 3: Raters in the study are the only raters of interest. All participants are measured by all raters.

In each of these three models, each rater may take one measurement on a participant, or each rater may take several measures on a participant. Standard notation for the different ICC types is "ICC(i,j)", where "i" refers to one of the models above, and "j" refers to the form of ICC (j=1 if the data have not been aggregated. If the data are aggregated, j is higher than one (e.g. 2, 3, 4, etc.) depending on how the data were aggregated). Shrout and Fleiss's guideline for choosing the appropriate ICC requires answering three questions. These are:

1. Did each rater/group of raters assess each participant?
2. Were the rater(s) the only rater(s) of interest, or were they selected randomly from a large population of raters?
3. Did the rater(s) measure each participant once or several times?

Answers to the first and second questions inform the suitable statistical model of ICC, represented by "i" in the ICC notation. The third question provides information on the form of ICC, which is represented by "j" in ICC notation. In the end, one will have the proper type of ICC (i,j) to assess reliability.

The answer to the first question determines whether one should use a one-way or two-way analysis of variance (ANOVA) for the analysis. In model 1, each participant is rated by a different set of raters randomly selected from a larger pool of raters. Since no participant is evaluated by more than one rater, the measurements are independent, and a one-way ANOVA is appropriate to analyze reliability (75). However, in models 2 and 3, there are repeated measures of the same participant by different raters. Therefore, the data are expected to be correlated (75). The appropriate analysis to account for this data dependence is a two-way ANOVA (75).

The answer to the second question helps to choose between models 2 and 3. In model 2, we have randomly selected a group of raters from a large population of raters, who rated all participants. We are interested in the reliability within this larger population of raters, and we must account for the increased uncertainty that occurs because we do not have measurements of every rater. In model 3, the raters in the study are the only raters we are interested in evaluating, and therefore there is one less source of error. However, generalizability to other raters is not valid (76). The formulas related to each model (75) are shown in Table 2.1.

Table 2-1. Formulas for three models of ICC

Model	Variance ratio	ICC formula
(Model 1) $x_{ij} = \mu + b_{j+} w_{ij}$	$\rho = \sigma_T^2/(\sigma_T^2 + \sigma_W^2)$	$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS}$,
(Model 2 and 3)	$\rho = \sigma_T^2/(\sigma_T^2 + \sigma_J^2 + \sigma_I^2 + \sigma_E^2)$	$ICC(2,1) = \frac{BMS-EMS}{BMS+(k-1)EMS+k(JMS-EMS)/n}$,
$x_{ij} = \mu + a_i + b_j + (ab)_{ij+} e_{ij}$	$\rho = \frac{\sigma_T^2 - \sigma_I^2/(k-1)}{\sigma_T^2 + \sigma_I^2 + \sigma_E^2}$	$ICC(3,1) = \frac{BMS - EMS}{BMS + (k-1)EMS}$.

n: number of participants, k: number of raters, BMS: Between-participants Mean Square, WMS: Within-participants Mean Square, JMS: Between-raters Mean Square, EMS: Error Mean Square.

In model 1, x_{ij} denotes the ith rating (i=1,...,k) on the jth (j=1,...,n) participant, μ is the overall population mean, b_j is the difference between ith participant's true score and μ of the jth participant's true score. Finally, w_{ij} (residual component) is the sum of the effects of the raters, the rater participant interaction, and error. Both b_j and w_{ij} are considered distributed normally with a mean of zero. Their variances are σT^2 and σW^2, respectively (75). In models 2 and 3, all the same k raters rate all n participants, a_i is the difference between μ and the mean of the i_{th} raters's measurements, $(ab)_{ij}$ is the degree to which the ith rater deviates from their own usual rating (tendencies confronted by the jih participant), and e_{ij} is the random error of the ith rater's measurement of the jth participant. The assumption is that a_i and e_{ij} are normally distributed with a mean of zero and variance of σJ^2 and σE^2, respectively. Finally, σI^2 is the variance due to interaction, $(ab)_{ij,}$ in models 2 and 3 (75). This interaction shows the degree to which the ith rater deviates from his own usual rating when measures the jth participant (75).

The third question defines "j" in ICC$_{(i,j)}$, where j=1 if the data have not been aggregated. If the data are aggregated, j is higher than one (e.g. 2, 3, 4, etc.) depending on how the data were aggregated prior to running ICC. A reliability based on the mean of measurement will always be greater in magnitude than the reliability of the individual measurement (77), which should be considered in the estimation of ICC. An example of choosing the mean of measurements for a reliability study is when an individual rating is too unreliable. In this case, a number of measurements are used to form the mean for the study (75). Answering all three questions leads

to different types of ICC with a unique formula to estimate that type of ICC (75). In the context of vision testing, we are interested in identifying the reliability of vision tests when clinicians outside the study use them (excludes Model 3). Because each participant is measured by the same clinician, we are left with Model 2.

2.4.2 Limits of Agreement

While the ICC evaluates the linear association between scores of tests (78), the agreement is assessed by the absolute difference between the scores of tests (79). The most common method of evaluating agreement when the test is a continuous measure is to calculate the limits of agreement (80). This can be illustrated graphically through the Bland-Altman plot (80).

Bland and Altman suggest constructing a scatter plot with the difference in the two results (A-B) against the mean value [(A+B)/2] (81). Bland and Altman note that plotting the difference against either A or B is misleading because it will always show an autocorrelation even when there is no relationship between A and B (82). As a simple example, consider A and B are randomly generated (no correlation), and we plot the difference (A-B) against B. As B is increased, A-B must decrease, and there will be a negative correlation (82). To visualize this, Table 2 includes randomly generated values (range 20 to 80) for a baseline test for 16 participants, randomly generated values for a test conducted one year later, the difference between the two values and the mean of the two values.

Table 2-2. Randomly generated test results for 16 people, including difference and mean of the two tests results

id	Baseline_Test	1-yr_Test	difference	mean
1	48	64	-16	56
2	34	22	12	28
3	38	21	17	29.5
4	65	41	24	53
5	73	67	6	70
6	71	56	15	63.5
7	38	38	0	38
8	26	51	-25	38.5
9	47	52	-5	49.5
10	24	75	-51	49.5
11	43	42	1	42.5
12	56	65	-9	60.5
13	76	70	6	73
14	80	30	50	55
15	79	46	33	62.5
16	30	20	10	25

Figure 2.1A represents the Bland-Altman plot, where the differences between the two values are plotted against the mean. As these two sets of numbers were randomly generated, there should be no correlation, and the observed slope of the regression line is close to 0. Figure 2.1B is a plot of the difference between the two values against one of the values. The slope of the regression line is clearly negative, as explained above.

A B

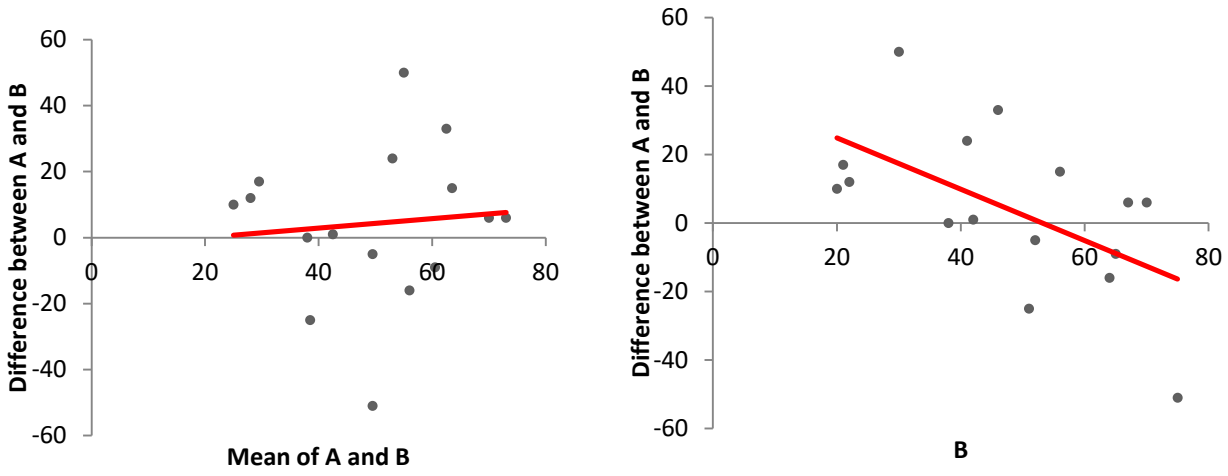


Figure 2.1. Comparing the regression line when the difference is plotted against mean and against one of the values; In A, the difference in test values is plotted against the mean, as suggested by Bland and Altman. The red line represents the regression line through the observed data. The slope is close to 0 as expected for randomly generated numbers that are not correlated. In B, the difference in values is plotted against one of the values. The red regression line is clearly negative, indicating a correlation between the two sets of values, even though they were both randomly generated.

The plot suggested by Bland and Altman for Limits of Agreement adds additional information (Figure 2.2). First, they superimpose the line of best fit using linear regression. Next, if the agreement is similar across all values of the test, the scatter of differences (e.g. SD at each test value) would be uniform. This is called the homoscedasticity. Because the mean ± 1.96 standard deviation includes 95% of normally distributed data, the 95% LoA is defined as the range from mean+1.96*SD to mean-1.96*SD (95% LoA). Bland and Altman suggest plotting the 95% LoA on the scatter plot to obtain a visual representation (79), and represents the range where we can expect ”individuals” test-retest values to fall 95% of the time (79). By using scatter plots, one can easily see any outliers or relationship between variance in measures and

size of the mean. The 95% LoA also provide a range of error that may be clinically acceptable, although this needs to be interpreted regarding the range of measures in the raw data (70).

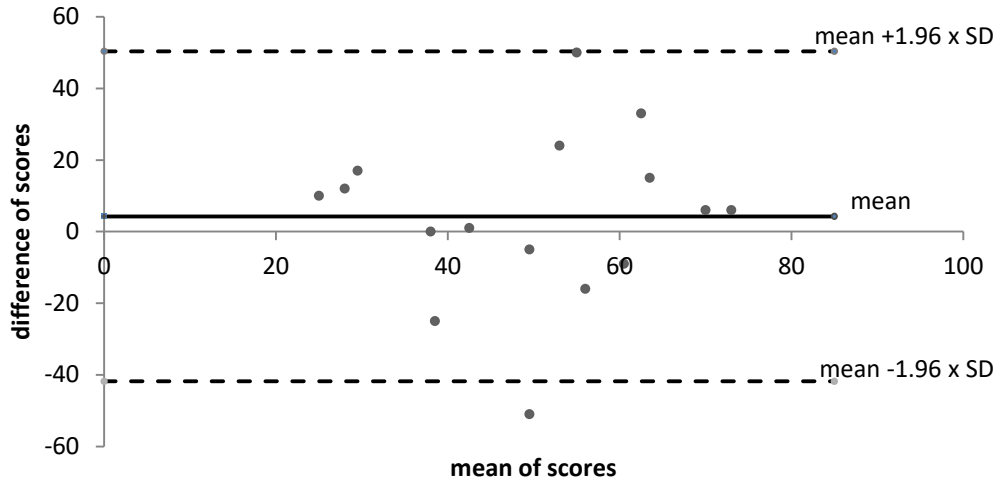


Figure 2.2. Bland-Altman plot for the Limits of Agreement of the data in table 2.1 The solid black line is the mean difference, and the two dash lines represent the mean difference ± 1.96*SD.

When the agreement of the test varies with the value of the test, the scatter of points around the mean will differ depending on the value of the test. For example, one would expect less error if one measures a length of 1m compared to measuring a distance of 100m. On the Bland-Altman plot, one would see a small SD at small distances and a higher SD at long distances, producing a funnel shape (Figure 2.3A). This is called the heteroscedasticity. In this case, we need to transform the data so that it is normally distributed. Often, this is achieved with a log-transformation of the data (Figure 2.3B), which is equivalent to analyzing/plotting the

percentage difference of the measurements (78, 79). The 95% LoA is again estimated as the mean difference ± 1.96*SD of the percentage differences.

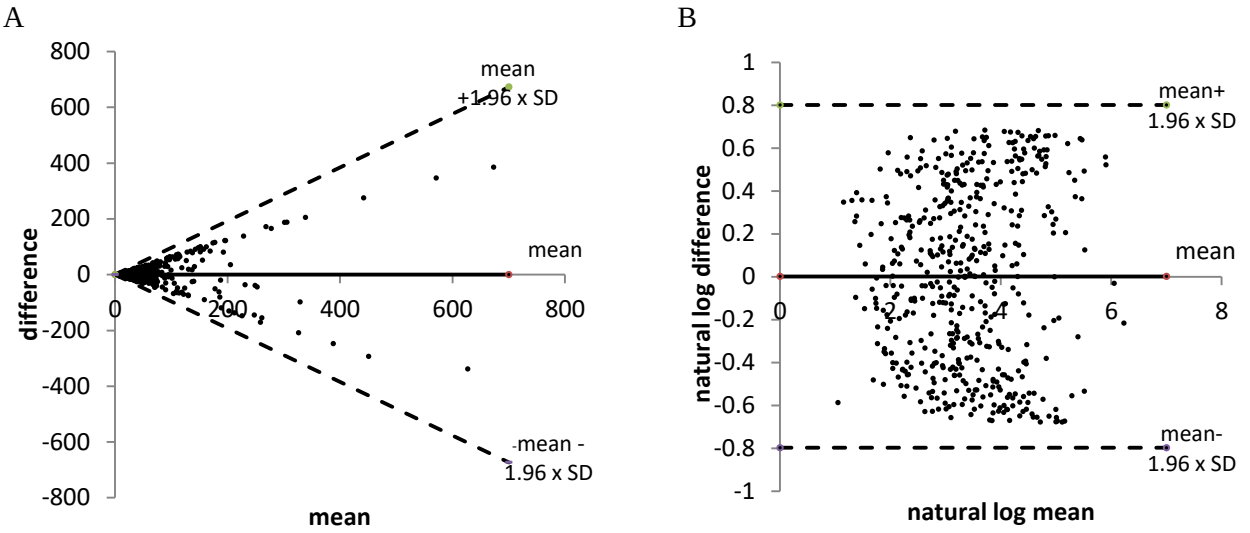


Figure 2.3. Bland-Altman plot using raw data and natural log-transformation.

In A**,** the Bland-Altman plot using raw data indicates the agreement of the test varies with the value of the test. In B, the Bland-Altman plot is plotted using natural log-transformed data. With log-transformed data, the axes represent the mean and difference of the natural log transformations.

In this section, I discussed general issues of reliability and agreement as measures of error (noise), which is essential if one wants to know if the change recorded in a test is larger than what is expected from the noise of the test. In the next section, I provide an overview of reliability characteristics for tests used to evaluate ten different aspects of visual function in patients with a concussion in this thesis. I provide detailed descriptions of how to conduct these tests in the methods section.

2.5 Overview of Vision Testing and Concussion

The reliability results for tests used to assess the different types of the visual function used in this study are summarized in Table 2.3.

Table 2-3. Summary of test-retest reliability studies of the ten vision test

Visual Function (reference #)	Test	Interval	Population	ICC	LoA
Gross Stereoscopic Acuity (25)	Randot® Stereotest	1 week	young healthy adults	0.86	± 55%
Gross Stereoscopic Acuity **(27)**	Titmus stereo fly and Frisby stereo tests	1 week	healthy preschool children	1.0	-
NPC **(83)**	Maples method	≤1 day	concussed athletes aged 9–24y	0.95 to 0.98[1]	-
NPC (23)	Prince rule	2-3 days	healthy young adults	0.65[2]	-
NPC (25)	Maples method	1 week	young healthy adults	0.54	±57.9
NPCb (24)	Astron International Accommodative Rule	1 week	healthy school children	0.92 and 0.89 [3]	-
NPCb (26)	RAF rule	≤ 1 week	adult aged 18 to 65y with idiopathic neck pain	0.84	-
NPCb (25)	Maples method	1 week	young healthy adults	0.64	±65.2
Positive Fusional Vergence, at distance (23)	Used a Risley prism	2-3 days	healthy young adults	0.72	-
Positive Fusional Vergence, at 3m (25)	used a horizontal prism bar	1 week	young healthy adults	0.49	±69.8
Positive Fusional Vergence, at 30cm (25)	used a horizontal prism bar	1 week	young healthy adults	0.54	±69.5
Positive Fusional Vergence, at 30cm (24)	Von Graefe	1 week	healthy school children	0.59 and 0.53 [2]	-
Negative Fusional Vergence, at 30cm (25)	used a horizontal prism bar	1 week	young healthy adults	0.66	±63
Negative Fusional Vergence, at 3m (25)	used a horizontal prism bar	1 week	young healthy adults	0.43	±68.8
Phoria, at 30cm (25)	PEDIG[4] protocol	1 week	young healthy adults	0.69	±122.0
Phoria, at 3m (25)	PEDIG protocol	1 week	young healthy adults	0.88	±123.6
Saccades (28)	prosaccade gain	Avg. 53 days	healthy participants	0.59	-
Saccades (25)	unpublished proprietary algorithm	1week	young healthy adults	0.34	±34%

[1] NPC test was performed 3 times consecutively in one day in concussed athletes.
[2] ICC was calculated by Rouse et al. (24) for data from Brozek et al. study (23).
[3] In this study two examiners rated the vision function and ICC reported for them separately.
[4] PEDIG: Pediatric Eye Disease Investigator Group

In the following, I review each test.

2.5.1 Gross Stereoscopic Acuity

A successful combination of two separate images from two eyes into one image in the brain is vital for stereopsis or 3D vision. The precision of stereopsis is essential for depth perception (84). Gross Stereoscopic Acuity is the smallest detectable separation in depth between two stimuli (85). It is quantified by assessing the ability to see images formed of dots that are displaced. (85).

Long et al. reported good reliability in Gross Stereoscopic Acuity in non-athlete young adults (ICC=0.86; 95% LoA = ± 55%) in a one-week test-retest study (25). They tested the ability to perceive depth with the Randot® Stereotest (Stereo Optical Co., Inc., Chicago, IL), in arc seconds (86, 87), the same method that we used in this thesis. Another study using Titmus stereo fly and Frisby stereo tests in children revealed excellent reliability (ICC=1.0) for Gross Stereoscopic Acuity (27). In addition, one study reported that 82.0% of their participants had identical results when the test-retest interval was within the same day in 100 healthy adults and children (19). Test-retest reliability of Gross Stereoscopic Acuity has not been determined for a one-year period.

2.5.2 Near Point of Convergence (NPC)

The near point of convergence (NPC) assesses the symmetric convergence ability of eyes when looking at a moving object. It measures the closest point in front of the patient's face in the space in the median plane when one eye deviates, looking at an object sitting at that point (88). (89). The RAF rule (RAF Binocular Gauge, Clement Clarke, Essex, UK) applies sophisticated equipment that allows more precise measuring of the distance in NPC (90).

Reliability studies have used a variety of methods, including Maples method, which is a non-accommodative test (29), or an accommodative target, such as Royal Airforce (RAF) rule (91) or Astron International Accommodative Rule (83). The one-week reliability study by Long et al. suggested moderate reliability for the Maples method of this test (ICC=0.54, 95% LoA= ±57.9) (25), which is the same method that we used in this study. Rouse et al. (24) reported a similar ICC of 0.65 for NPC in healthy adults based on the data from Brozek et al. (23). This was a study on NPC on six occasions in six subjects, using Prince rule, an accommodative target in 2-3 days (23). It was not clear from the report whether single or multiple examiners assessed the test. Another study reported excellent test-retest reliability (ICC= 0.95 to 0.98) in concussed athletes, aged 9-24 years, in three consecutive sessions in one day (83). No published studies evaluated the one-year test-retest reliability of NPC.

2.5.3 Near Point of Convergence break (NPCb)

This test evaluates the symmetric convergence ability of eyes when looking at a moving object, using the same methods as NPC, but the test ends when the participant has subjective double vision due to the inability of the eyes to converge (92). The score of the test is the distance between the bridge of the nose and the point (in cm) where double vision occurs. One-week test-retest reliability by Long et al. defined ICC=0.64 for NPCb and a 95% LoA of ±65.2 (25). Giffard et al. found that the RAF rule NPCb had good test-retest reliability (ICC=0.84) "within one-week" among adults aged 18 to 65 with idiopathic neck pain (26). However, it was unclear if they assessed participants at a one-week interval or less than one week. In Rouse et al. study, two examiners measured NPCb one week apart in fifth and sixth-grade students using the Astron International Accommodative Rule. Test-retest reliability (intra-rater reliability) of the NPCb was excellent for one examiner and good for the other examiner (ICC= 0.92 and 0.89)

(93). Various methods have been used to define NPCb, such as Maples method (29), which is a non-accommodative test, and RAF rule (91), which uses an accommodative target. One-year test-retest reliability of NPCb has not been studied.

2.5.4 Positive Fusional Vergence

This test examines how well a participant can converge the eyes in order to fixate light on their retina at a near distance (30cm) and far distance (3m), measured in prism diopters. This test assesses binocular fusion ability with convergence, which means the participant must move both eyes medially to focus on the object. In one-week test-retest reliability study by Long et al. Positive Fusional Vergence showed moderate reliability at 30cm (ICC=0.54, 95% LoA= ±69.5) and poor at 3m (ICC=0.49, 95% LoA= ±69.8) (25). Brozek et al. examined 2-3 days test-retest reliability of Positive Fusional Vergence at distance in healthy adults and found an ICC of 0.72, indicating moderate reliability (23, 24), but it was not clear whether single or multiple examiners performed the assessments (23). Rouse et al. study showed moderate test-retest reliability (ICC= 0.59 and 0.53, reported for two raters separately) for Positive Fusional Vergence at 30cm among healthy children (24). There are no one-year test-retest reliability studies.

2.5.5 Negative Fusional Vergence

This test examines how well a participant can diverge the eyes in order to fixate light on their retina at a near distance (30cm) and far distance (3m), measured in prism diopters. The difference from Positive Fusional Vergence is that this test assesses binocular fusion ability with divergence, which means the participant must move both eyes laterally to focus on the object. Negative Fusional Vergence showed moderate (ICC=0.66, 95% LoA= ±63) and poor (ICC=0.43,

95% LoA=±68.8) reliability, at 30cm and 3m respectively in one-week test-retest study (25). There are no published one-year test-retest reliability studies.

2.5.6 Phoria

Phoria evaluates the natural deviation of the eyes at near and far targets. We can measure it in prism diopters, with a prism and alternate cover test using a target placed at 3m (far vision) and 30cm (near vision) (25). This is different from strabismus, a condition in which the eyes do not properly align with each other even when not focusing on an object (94). Therefore, phoria cannot be measured when there is strabismus. In Long et al. one-week test-retest study of young adults without strabismus, ICC for phoria was 0.88 (95% LoA=±123.6) at 3m and 0.69 (95% LoA=±122.0) at 30cm (25). There are no published one-year test-retest studies.

2.5.7 Saccades

Saccades are rapid eye movements that abruptly alter the point of fixation. The test for saccades assesses three functions: quality, synchronization and correction of rapid eye movements when trying to focus on flashes of light on a screen. Studies suggest SRC can affect various aspects of saccades, such as antisaccades, which is the initiation of a rapid eye movement in the opposite direction to a sudden visual target (95), or remembered saccades, in which the eyes move toward a remembered point without a visual stimulus (96). In addition, the examination of patients who had PCS for 3–5 months showed that they had worse performance on Saccades than those who had recovered from a concussion. One study found that patients who had recovered well from a concussion did not have saccadic dysfunction (97). Moganesware et al, estimated a moderate reliability for prosaccade gain in an average of 57 days interval in healthy participants (ICC= 0.59) (28). In the prosaccade, they asked participants' to follow a

target as quickly and accurately as possible. Prosaccade gain was calculated as a percentage of saccades amplitude divided by target amplitude multiplied by 100 (28).

There is one report on test-retest reliability of Saccades among healthy participants over a nineteen-month interval. However, the test components are very different from the test of Saccades that was performed at APEXK. There are no other published one-year test-retest reliability studies of Saccades. Although saccades are normally evaluated on three functions, each measured on an ordinal scale, in this thesis, we used an overall score provided by the participating clinician that was based on a proprietary algorithm. Long et al. (25) found the one-week test-retest reliability using this method to be poor (ICC=0.34). However, the 95% LoA for Saccades was the lowest (34%) among the ten vision tests they examined (25).

3 Material and Methods

3.1 Study design

The study described in this thesis is a historical cohort observational study. We carried out this study to identify the one-year test-retest reliability of ten vision tests in young Canadian athletes followed by Institut National du Sport du Quebec (INSQ). Some of these athletes had a yearly examination done by a sports medicine physician at INSQ and vision tests done by a clinician trained in orthoptic[4] testing, who was one of the industry partners at APEXK Inc (APEXK). Patient data were available in electronic charts of the athletes in INSQ and APEXK, where athletes had their medical follow-ups and vision tests done. This study was approved by the McGill University Faculty of Medicine Institutional Review Board.

3.2 Participants

The population for the study included Canadian athletes over 16 years of age, followed by the INSQ from 2015–2018. We only included athletes who had completed two baselines (preseason) annual evaluations within a 365-day (± 30 days) period. We excluded athletes who suffered a concussion between annual evaluations or had received orthoptic training between the baseline measures. The reason to exclude these athletes was that any concussion could potentially affect a test result, which would invalidate any reliability interpretation for "test-retest" results. In addition, any additional orthoptic training or testing in between test and retest

[4] Orthoptic training is training for allied health professionals. This training makes them competent in the diagnosis and non-surgical treatment of vision problems in eye movement, eye alignment and binocular vision. (https://www.internationalorthoptics.org/about-us/profile/professional-role/)

could have a similar consequence through the learning effect. If an athlete had more than two baseline testings, we only included the first two tests. We also excluded athletes with conditions that may affect binocular vision and Saccades. These included a history of strabismus, neurological disorders such as migraine, and those medically treated for depression, anxiety or other psychiatric condition. Data were collected from electronic medical charts of one clinician trained in orthoptic measures at APEXK and one sports medicine physician at INSQ.

3.3 Measures

Our goal in this test-retest study was to estimate the stability of the test results (reliability and agreement) in a one-year period. Therefore, we excluded any test-retest out of the one year ±30 days interval. We did not define a cut-off for a test result to distinguish normal and abnormal results. Therefore, in this study, we used the actual numerical results of the vision tests instead of the dichotomized (normal/abnormal) version of the test.

At the beginning of each season, the orthoptic-trained clinician at APEXK Inc assessed baseline testing of ten vision tests in all athletes. 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 follows.

3.3.1 Gross Stereoscopic Acuity

Gross stereoscopic acuity measures the ability to perceive depth. We used the Randot® Stereotest (Stereo Optical Co., Inc., Chicago, IL), measured in arc seconds. Seated participants

wearing polarized glasses were asked to hold the testing booklet 16 inches from their faces. 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 (86, 87).

3.3.2 Near Point of Convergence (NPC)

NPC assesses the ability to symmetrically converge the eyes when looking at a moving target (98). 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, in centimetres, between the bridge of the nose and the distance of the target at the closest point at which the individual could maintain 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 (91), but a more recent study suggested 5 cm should be considered the upper limit of normal values (99).

3.3.3 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 (92). The score of the test is the distance between the bridge of the nose and the point where double vision occurs (in cm), 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 (100); however, data on adults with normal vision suggest a breakpoint of approximately 5.0 to 7.5 cm (101).

3.3.4 Positive Fusional Vergence

This test examines how well an athlete can adapt to challenges in fixating light on their retina at a near distance (30cm) and far distance (3m), measured in prism dioptres (25). 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 dioptres (98, 102, 103).

3.3.5 Negative Fusional Vergence

This test is similar to 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 placed at near (30cm) and far (3m) (25). It is 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 dioptres (98, 102, 103).

3.3.6 Phoria

This test assesses the natural deviation of the eyes at near and far targets (25). It is measured in prism diopters, with a prism and alternate cover test using a target placed at 3m (far vision) and 30cm (near vision) from the participant. While the seated participant is fixating on the target, the clinician covers and uncovers each 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 is progressively increased until no shift in the eyes is seen. The score of the test is the rating of the prism that cancels the eye movements, with lower scores representing less phoria. There is no normative data identified for this test (25).

3.3.7 Saccades

This test examines the eyes' ability to perform saccadic movements, which are rapid eye movements that abruptly alter the point of fixation. In the APEXK clinician's version of this test, participants assume a tandem stance (heel-to-toe with a 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 (rates: bad, medium, good) and saccadic correction (rates: many corrections, few corrections, no corrections). The three sub-scores are 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 clinician conducting the test calculates an overall score based on a proprietary algorithm.

3.4 Analysis

We evaluated test-retest reliability and agreement, using the Intraclass Correlation Coefficient (ICC) and 95% limits of agreement (95% LoA). We used Shrout and Fleiss's guideline (104) to choose the appropriate type of ICC for the test-retest study. In our study, the same clinician examined all athletes, and we were interested in generalizing our results to other clinicians. Therefore, the appropriate ICC model was a 2-way ANOVA. Finally, we could determine the ICC model by knowing that we did not aggregate the data. Therefore, j was equal to one. At this point, we had the appropriate type of ICC, which was $ICC_{(2,1)}$. We used this formula to estimate ICC (75):

$$\mathrm{ICC}(2,1) = \frac{\mathrm{BMS} - \mathrm{EMS}}{\mathrm{BMS} + (\mathrm{k} — 1)\mathrm{EMS} + \mathrm{k}\,(\mathrm{JMS} - \mathrm{EMS})/\mathrm{n}}$$

BMS: between participants mean square; EMS: error mean square; k: number of raters/measurements; JMS: between raters mean square; n: number of participants.

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 (105). We report the LoA in the raw units of the scale used by clinicians at APEXK. To compare LoA across tests, we also standardized the scores and reported them as percent differences, [(T1- T2)/ mean(T1&T2)]*100 (81, 106). 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 were provided in parentheses in the graphs 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. Also, we report the mean (SD) for continuous variables at baseline.

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-tests) (107). 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 (108).

3.5 Data sources

We used two sources of data in this study. First, we used electronic medical charts of the chief medical officer of the Institut National du Sport du Quebec (INSQ) to obtain demographic data on athletes who were referred for vision testing. All INSQ athletes visit the same sports medicine physician at INSQ for their health-related conditions, including possible concussions. Second, we obtained data from vision testing for athletes from the electronic files of a clinician trained in orthoptics at APEXK Inc. The sports medicine physician at INSQ referred athletes to APEXK for vision tests for baseline testing and after a concussion. Athletes may also go to APEXK for vision training without a referral.

3.5.1 INSQ data

Data were extracted from the INSQ database on athlete visits for concussions in an Excel sheet. We renamed variables to include “INSQ” for clarity and consistency throughout this document. In the Excel sheet, each row represented one athlete, and each column represented the

athlete's values for one of these variables: a unique ID (INSQ_ID), sex (F or M; INSQ_sex), date of birth (INSQ_DOB), and type of sport. Additional columns provided dates for each concussion (INSQ_concussion_1, INSQ_concussion_2, etc), and the date the athlete was referred for vision testing (INSQ_session_1, INSQ_session_2, etc).

3.5.2 APEXK data

The vision data were provided from the company APEXK in an Excel sheet, where each row represents one testing session. For example, there would be five rows if the same athlete had vision tests on five different days for the same concussion. Although we requested APEXK extract only INSQ athlete data from its database, some none-INSQ patients, such as patients referred from other physicians, could mistakenly be included. We renamed variables to include "APEXK" for clarity and consistency throughout this document. For each patient visit, the clinician recorded:

- unique ID (APEXK_ID)
- self-reported age
- sex
- date of concussion (APEXK_concussion)
- date the vision test was conducted (APEXK_session)
- symptoms before vision testing
- vision test scores
- session number
- the objective of the session; The four possible objectives of a testing session at APEXK (APEXK_session) were 1) testing before the sports season (baseline), 2) testing

following a concussion with or without visual training (concussion), 3) vision training to improve visual abilities and performance in non-concussed athletes (training) and 4) vision testing sometimes occurred long after vision training to determine if the visual function had deteriorated (follow-up).

In the APEXK data, a consecutive series of training sessions for the same concussion were related. APEXK numbered each session according to the following system. Sessions were numbered sequentially, so that session 1 occurred before session 2. Unrelated sessions, such as two different sequential baseline tests, would have different integer numbers (e.g. sessions 1 and 2). Related testing sessions were given the same integer number (e.g. 3) but different suffixes (e.g. sessions 3, 3.1 and 3.2 for three testing sessions related to the same concussion). Related sessions could all be for the same concussion or related to a series of visits to evaluate the effects of daily training to enhance performance. The first session for any group of associated sessions was always an integer (e.g. 1, 2, etc.).

3.6 The logic for tidying the data

Our goals in tidying data were to ensure the data from both INSQ and APEXK data were valid and to identify athletes who had two baseline vision tests within 365 ± 30 days, without any concussion, testing or training in between the baseline tests. To do this, we started with the APEXK data as our source file because it was supposed to contain all the vision test results for all INSQ athletes. We then made corrections where necessary. We accomplished this process in close collaboration with the sports medicine physician at INSQ and the clinician at APEXK.

3.6.1 Merging data from both sources

3.6.1.1 Gathering and validating IDs

Our first objective was to ensure that we had a complete list of all INSQ athletes who had vision test scores in the APEXK data. Our objective was to exclude APEXK data that did not come from INSQ athletes and to ensure all INSQ athletes that had vision tests were in our data.

APEXK data had one field for all athletes that identified where the individual came from, among all patient referral sources. All athletes from INSQ were supposed to have the same code in this field. If a patient was not an INSQ athlete, the code in this field was supposed to be different. To gather and validate the eligible IDs, we obtained a list of athletes' IDs from the APEXK (APEXK_ID) and INSQ (INSQ_ID). Then, we followed the process below to identify all athletes seen by the sports medicine physician and had results of their vision tests from APEXK.

3.6.1.1.1 Comparing the list of APEXK_ID with the list of INSQ_ID

First, we copied APEXK_IDs and pasted them into a new Excel file, named "temporary file". Next, we copied and pasted INSQ_IDs into a new column named INSQ_IDs in the temporary file. At this point, we had all IDs reported from both data sources in the temporary file. If an athlete's ID was among both the INSQ_IDs and APEXK_IDs, they were considered potentially eligible for our study. As these IDs already existed in both files, we deleted them from the temporary file, leaving only IDs that were either only in INSQ or only in APEXK data. These IDs that did not match needed to be further explored.

3.6.1.1.2 ID in INSQ_IDs but not in APEXK_IDs

First, APEXK searched through their data to identify additional records that were omitted in their first exported data. In some cases, the errors were due to minor errors in the ID numbers. In other cases, an INSQ athlete was miscoded as a non-INSQ athlete. These records were retrieved on a subsequent export of the data and added to the APEXK data.

If the INSQ_IDs were not found in the APEXK database, they were sent back to the sports medicine physician at INSQ to confirm the validity of the IDs provided. If the sports medicine physician found that the INSQ_ID contained a mistake, the INSQ_ID was corrected. Then this corrected INSQ_ID was checked against APEXK_IDs in the temporary file for a matching ID. If no matched ID was found in the APEXK data, this corrected INSQ_ID was sent to APEXK to verify that the athlete did not exist in the APEXK data and never had the test. If APEXK confirmed that there was no mistake in the registration of data at APEXK and the individual was not tested at APEXK, we deleted this ID from the temporary file. We did not add this ID to the APEXK data. An INSQ athlete would not have APEXK data if the athlete was never referred for vision testing at APEXK, or the athlete was referred but had not gone for the testing. All INSQ_IDs without any APEXK data were deleted from the temporary file, leaving only APEXK_IDs that did not match any INSQ_ID.

3.6.1.1.3 ID in APEXK_IDs but not in INSQ_IDs

If an ID was in APEXK_IDs but not INSQ_IDs, APEXK sent the sports physician at INSQ all the relevant data, including the athletes' names. This information was not available to the researchers. The sports medicine physician then reviewed the records to confirm whether the athlete was an INSQ athlete or not. When INSQ could not identify a name among its athletes, the

APEXK data was considered to be in error, and the corresponding APEXK_ID was deleted from the APEXK data file containing the vision test scores. If the names were identified in the INSQ data, the INSQ_ID was corrected, and the additional relevant data were provided by INSQ in subsequent data extraction.

We followed this process to validate all IDs in the temporary file. At the end of this step, we identified all athletes who were seen by the INSQ sports medicine physician and were tested for their vision at APEXK. In the next steps, we used the APEXK data sheet with all vision testing results as the base file and corrected/added relevant information based on the INSQ data.

3.6.1.2 Gathering and validating date of birth (DOB) and sex

The INSQ extracted data on DOB and sex from the Quebec medicare numbers of athletes. In the APEXK data, these were self-reported. We did not have access to the Quebec medicare information for confidentiality reasons. Therefore, we considered the INSQ data to be more reliable. We deleted self-reported DOB and sex in the APEXK data and added the corresponding DOB and sex from the INSQ data based on matching INSQ_ID and APEXK_ID. We verified that DOB was present for all records, and all were within the expected ranges for INSQ athletes. We verified that sex was present for all records, and always equal to F or M.

3.6.2 Standardizing date format

At this point, we had gathered validated APEXK_IDs, INSQ_DOB and INSQ_sex in the APEXK data. Our analysis required merging dates from the INSQ and APEXK. The APEXK_session dates (testing dates) were created through an automatic time-stamp and were

considered correct. We standardized the format of the dates, so it was the same in both data sources.

3.6.3 Validating concussion dates

In the APEXK data, we had self-reported concussion dates (APEXK_concussion), and we were not sure about their accuracy. The APEXK data was in a "long" format, with one row for each vision testing session for each athlete. In INSQ data, the sports medicine physician recorded INSQ_concussion date when diagnosed. These data were in "wide" format, with one row for each athlete and a different column for each concussion date (INSQ_concussion_1, INSQ_concussion_2, etc.).

We verified that all relevant concussion dates in the INSQ data (INSQ_concussion_1, INSQ_concussion_2, etc.) were included in the APEXK data and vice versa. This important step allowed us to identify athletes who are not eligible for the study in this thesis. Eligibility required that athletes have two baseline tests with no concussion or training occurring in between the tests. If there was a discrepancy, we reviewed the files with APEXK and INSQ. If there remained a concussion date in the INSQ data that was not included in the APEXK data because the athlete was never seen by APEXK, we inserted a row in the APEXK data so that we could identify a concussion that occurred between two baseline tests (i.e. ineligible for our study). There were other possibilities, such as the presence of a concussion date in the APEXK data but not in the INSQ data or a minor discrepancy between INSQ data and APEXK data in the concussion date, such as concussion dates less than two weeks different. We addressed these in the next steps.

We sorted the APEXK data, which could have many rows for each athlete, by APEXK_ID and then the session date. We also sorted the INSQ data, which has one row per athlete, by INSQ_ID. Elementary checks (e.g. concussion date is after the date of birth) were corrected. Between these data sets, it is possible that 1) an INSQ concussion date matched an APEXK concussion date; 2) there was a concussion date in the APEXK data but not in the INSQ data; 3) there was a concussion date in the INSQ data but not in the APEXK data; or 4) there were concussion dates in both INSQ and APEXK data, but they did not match. We added one new column to the APEXK data named final_concussion_date to record the correct date after the resolution of possible discrepancy on this date. If the dates from both sources matched, this date was entered as the final_concussion_date. When there was a discrepancy, we followed the steps below, depending on the discrepancy identified.

For all steps, we began by looking at the first INSQ_ID and any concussion dates for this ID. It could be possible that there was no concussion or at least one concussion for this athlete in INSQ data.

3.6.3.1 No concussion in INSQ data

If there was no concussion for the first INSQ_ID, we looked at the corresponding APEXK_ID in the APEXK data to see whether there were any entries in the APEXK_concussion field. We sequentially looked at each of the rows in the APEXK data for the same APEXK_ID until the last row. If there was no concussion date in any of the APEXK data for this APEXK_ID, this meant both data sources concurred no concussion happened during the study period, for that athlete. In this case, we left final_concussion_date blank and moved to the next INSQ_ID.

When there was no concussion date in the INSQ data and at least one concussion date in the APEXK data for that individual, we checked whether the concussion dates were within the eligibility period of the study, 2015-2018. If any of the rows contained APEXK_concussion dates outside of this period, we left the corresponding final_concussion_date for that row blank (excluded the dates). We did not delete the row because it would remove important data we needed for further steps that determine eligibility for our study. If any of these APEXK_concussion dates were within the eligibility period, we sent the relevant information to INSQ for verification. If INSQ confirmed a concussion even though the original data export did not include that information, we corrected the INSQ data and added the INSQ concussion date to the appropriate row in the APEXK_concussion column. When INSQ had no record of a concussion, we considered INSQ data more reliable and left the final_concussion_date blank on the corresponding row in the APEXK data. Our rationale was that an athlete did not have access to the vision tests for a concussion without the INSQ physician referral, and APEXK_concussion dates were self-reported by the athlete and not routinely verified.

3.6.3.2 One or more concussion dates in INSQ data

If there was at least one concussion date for an ID in the INSQ data, we started with the first concussion date for this ID. If this date was not within the eligibility period of the study (2015-2018), we deleted it and proceeded to the next concussion date for that athlete. If this date was within the eligibility period, we checked whether APEXK_concussion dates for the same ID were within the eligibility period. If they were also within the eligibility period of the study, we compared the INSQ concussion date against available APEXK_concussion dates for the same athlete. If we found a matching concussion date in the APEXK data, we entered it into the final_concussion_date column in the APEKX data on the same row that contained the matching

APEXK_concussion. If we did not find a matching APEXK_concussion date, but any APEXK_concussion dates were within two weeks of the INSQ_concussion date, we accepted INSQ_concussion date as the final_concussion_date for the corresponding row. Our rationale was that an athlete might forget the exact time of a concussion, and we can accept INSQ_concussion date as a more accurate date given they were close to each other. Our choice to accept INSQ dates as valid if the difference in dates was only two weeks was based on expected errors due to memory recall but remain somewhat arbitrary. If APEXK_concussion dates and INSQ-concussion dates were more than two weeks apart, we were concerned they might represent two different concussions, and we verified the data with both INSQ and APEXK.

If there was no possible corresponding concussion date in the APEXK data available, we asked INSQ and APEXK to verify their data for this athlete. If there was still a discrepancy after verification, we considered the INSQ data indicating a concussion occurred more accurate for the reasons mentioned previously. In these cases, we inserted a row in the APEXK data immediately before the first date of APEXK_session that occurred after the INSQ_concussion date. We included APEXK_ID, DOB, and sex. We then inserted the INSQ_concussion date into the final_concussion_date column in the APEXK data and left all other fields blank. We repeated this process for all other concussion dates of each INSQ_IDs. We repeated the entire process for all validated APEXK_IDs, available in the APEXK datasheet.

At this point, our data file identifies all INSQ athletes who went for vision tests. Each row includes an ID, DOB, sex, and a validated concussion date. Next, we had to identify the four possible objectives of a testing session, which were:

- Baseline: testing before the sports season
- Concussion: testing following a concussion with or without visual training
- Training: vision training to improve visual abilities and performance in non-concussed athletes
- Follow-up: vision testing occurred long after vision training

3.6.4 Validating APEXK_session objectives

Although the APEXK data included information on session objectives, we were not sure about the accuracy of these data. These data were defined based on the presence or absence of self-reported concussions. We added one new column to the APEXK data named session_objective_baseline to register a validated objective for each session in APEXK, which could be different from the APEXK_session objective. If a validated session objective was baseline testing, we considered 1 for the value for this column, otherwise 0.

As explained in Section 1.5.2, the sessions in the APEXK data were numbered sequentially. Unrelated sessions, such as two different sequential baseline tests, would have different integer numbers (e.g. sessions 1 and 2). Related sessions, such as three visits for the same concussion, were given the same integer number (e.g. 3) but different numbers after a point (e.g. sessions 3, 3.1 and 3.2 for three testing sessions related to the same concussion). The first session for any group of associated sessions had only one level of number and was always an integer (e.g. 1, 2, etc.).

We started with the first row of the first APEXK_ID and reviewed the session number and APEXK_session. The session number in the first row of any related session should begin with an

integer. If this was not the case, we asked APEXK to verify what data were missing. After verification, we corrected the data accordingly and followed the next step.

If the session number of a row had only one integer (one level of numbering without a point) and there was one final_concussion_date within two months before the APEXK_session date of that row, we considered the session objective to be concussion testing. In this case, we entered 0 in the session_objective_baseline column of the corresponding row. Otherwise, we considered the session objective baseline testing.

If the session number had two levels, separated by one point (e.g. 3.2) and the previous rows with the same level session number (e.g. 3, 3.1) were not associated with a concussion, we accepted the session objective as training because it was related to a previous session. Therefore the objective could not be to obtain a baseline measure. We entered 0 in the session_objective_baseline column of the corresponding row. The first session of these related sessions (identified by the integer without a point) was still considered equivalent to a baseline session because it was not related to concussion and had not had any training prior to it.

If the session number consists of three levels of numbers, separated by two points (e.g. 3.3.1, 5.4.1), we considered the session objective as a follow up. A follow up was done after training to determine if the visual function had deteriorated. An example of a follow up would be a session with a session number of 3.3.1 in this series: 3, 3.1, 3.2, 3.3, 3.3.1. In this case, we assigned 0 value in the session_objective_baseline column of the corresponding row, indicating that this is not a baseline test. We repeated the whole process for all other APEXK_IDs.

At the end of this step, we had all data entry corrected in the APEXK datasheet, including IDs, DOBs, sexes, concussion dates, session dates, and baseline session objectives. Moreover, we formatted the data and eliminated ineligible athletes.

3.6.5 Defining eligible baseline tests

The objective of this step was to extract the data that meets the inclusion criteria for our study in this thesis, i.e. data related to athletes with two baseline tests within 365 ± 30 days of each other with no concussion, testing or training in between for APEXK_IDs. We deleted all other data for these athletes at the end of this step.

We sorted the validated APEXK data based on the APEXK_ID and then APEXK_session. If there were less than two rows for the same APEXK_ID where session_objective_baseline = 1, this APEXK_ID had <2 baseline tests and was not eligible for our study. We deleted all data related to this APEXK_ID. If there were at least two rows where session_objective_baseline=1, the patient had ≥2 baseline tests. Then we checked whether the corresponding APEXK_session dates had any intervening concussion, training or testing. If the rows with session_objective_baseline=1 were not on two consecutive rows, this means that testing, training or concussion had occurred between the two baseline tests. Therefore, these baselines tests were not eligible for the study.

Next, we verified the number of days between baseline vision tests in INSQ athletes who met all other criteria for eligibility. If the APEXK_session dates were not within 365 ± 30 days of each other, the data were not eligible for our study. When there were three consecutive baseline tests without any intervening concussion or training, we checked the difference between the second and third baseline tests to determine if they were within 365 ± 30 days of each other.

In these cases, we did not test non-consecutive baseline tests (e.g. baseline test 1 against baseline test 3) because even one baseline test (i.e. the 2nd baseline test) could be considered "training" and affect the result of a consecutive baseline test. For each APEXK_ID, we continued this process for all consecutive baseline test rows (APEXK_session_dates=1). Where we found eligible baseline tests, we deleted all other rows corresponding to the APEXK_ID.

We followed the whole process for all APEXK_IDs. Although it was theoretically possible that we would observe more than two eligible baseline tests for athletes, this did not occur. We had decided we would just consider the first two eligible dates for each athlete, and we would ignore the rest of the baseline tests if they were present. This is because, under the assumption of $ICC_{(2,1)}$, each participant is assessed once for the reliability analysis. In the end, we had two rows for each eligible APEXK_ID, indicating two baseline tests without a concussion, testing or training in between. This Excel file was the import data source for the analysis.

4 Results

4.1 Preface:

The result of this thesis has been published in F1000 Research Journal, as:

Aloosh M, Leclerc S, Long S, Zhong G, Brophy J, Schuster T, Steele R, Shrier I. One-year test-retest reliability of ten vision tests in Canadian athletes [version 3; peer review: 1 approved]. F1000Research 2020, 8:1032

(https://doi.org/10.12688/f1000research.19587.3)

4.2 Manuscript: One-year test-retest reliability of ten vision tests in Canadian athletes

The manuscript published in F1000 Research

Open for Science

Published online:

https://doi.org/10.12688/f1000research.19587.2

Mehdi Aloosh, MD

One-year test-retest reliability of ten vision tests in Canadian athletes

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

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 0.21 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 ±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.

Keyword:

concussion, vision tests, binocular, saccades, reliability

Corresponding author: Ian Shrier

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.

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 [9–17, 39-41]. 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 [9–17, 39-41], except for one test of saccades[44]. 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 [18]. 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 day 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 4.1. 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 [19]. 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 [20]. 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 saccades score, based on an unpublished proprietary algorithm developed by the clinician who performed the testing.

Table 4-1. 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 [21–23].

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 [21–23].
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" [22], 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 [24] ,but a more recent study suggested 5 cm should be considered the upper limit of normal values [25] .
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 [26] ; however, data on adults with normal vision suggest a breakpoint of approximately 5.0 to 7.5 cm [27] .
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 [28, 29] .
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 subscores 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) [30] and 95% limits of agreement (LoA) [31] . 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 [32] . 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 [31, 33] . 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) [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 3.4.3 [35]. 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 4.1). 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.

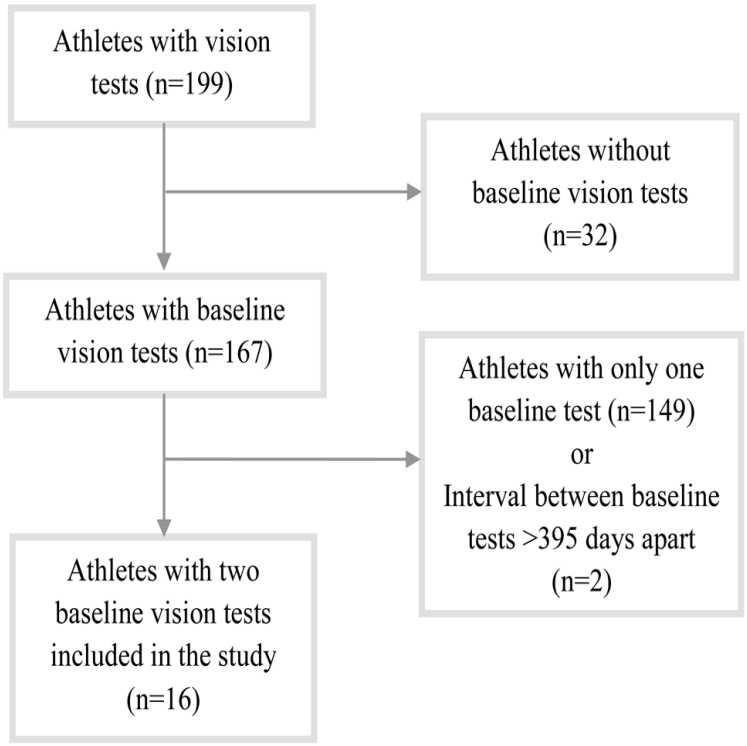


Figure 4.1. Patient flow diagram.

The range of scores observed for each vision test can be found in each of the reliability figures (Figure 4.2– Figure 4.4) [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 4.2). 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 conducted a sensitivity analysis excluding the outlier. This decreased the ICC from 0.93 to 0.53, and increased the LoA to ±43.5%.

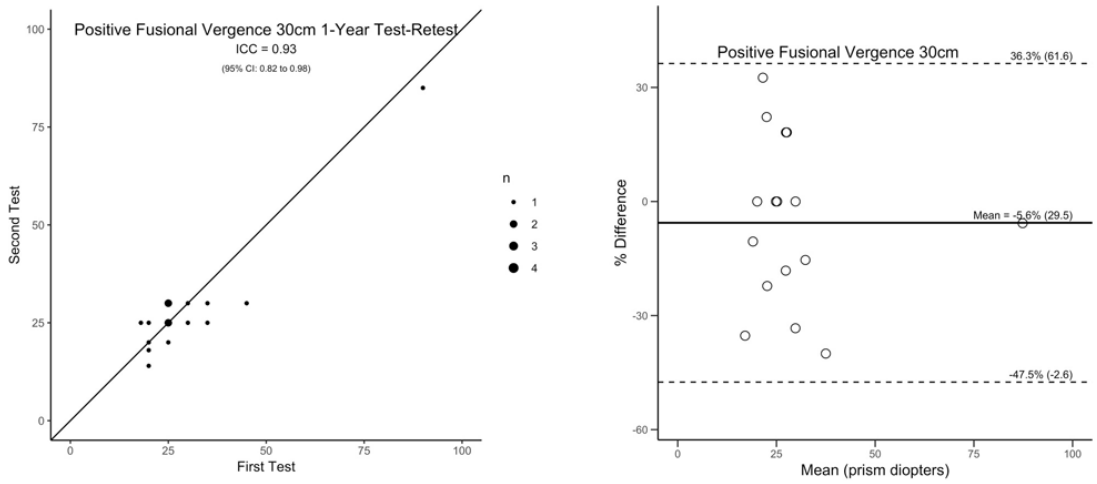
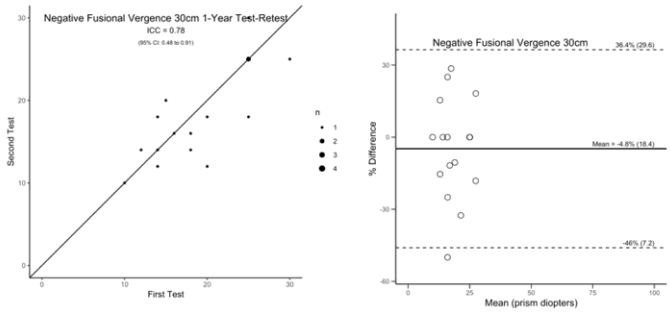
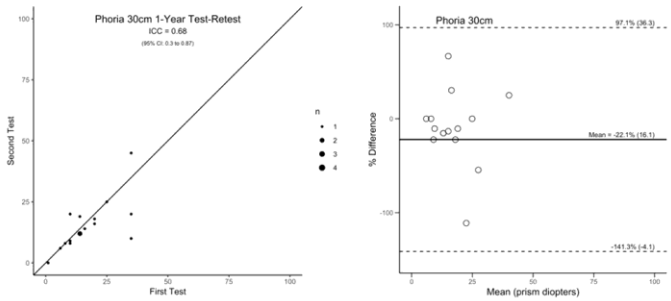


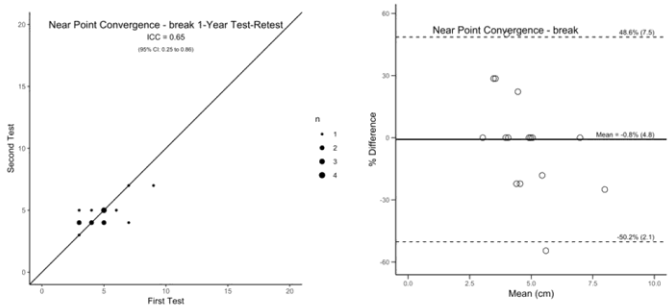
Figure 4.2. Vision test with excellent one-year test-retest reliability.

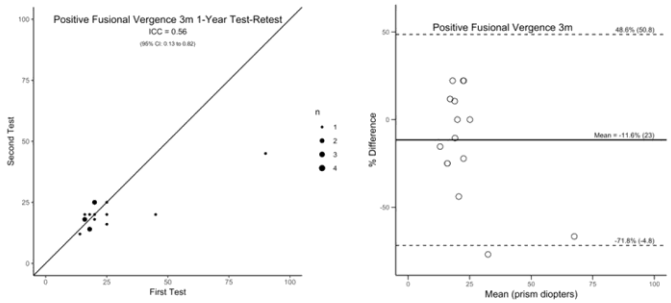
(**A**) Scatter plot of test-retest reliability for Positive Fusional Vergence at 30cm. Identity line represents perfect agreement between the test-retest values; ICC refers to the Intraclass correlation coefficient and 95%CI refers to the 95% Confidence Interval. "n (1,2,3,4)" refers to the number of 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.


Negative Fusional Vergence 30cm 1-Year Test-Retest
ICC = 0.78
(95% CI: 0.48 to 0.91)
Second Test
First Test
n
1
2
3
4
Negative Fusional Vergence 30cm
36.4% (29.6)
Mean = -4.8% (18.4)
-46% (7.2)
% Difference
Mean (prism diopters)


Phoria 30cm 1-Year Test-Retest
ICC = 0.68
(95% CI: 0.3 to 0.87)
Second Test
First Test
n
1
2
3
4
Phoria 30cm
97.1% (36.3)
Mean = -22.1% (16.1)
-141.3% (-4.1)
% Difference
Mean (prism diopters)


Near Point Convergence - break 1-Year Test-Retest
ICC = 0.65
(95% CI: 0.25 to 0.86)
Second Test
First Test
n
1
2
3
4
Near Point Convergence - break
48.6% (7.5)
Mean = -0.8% (4.8)
-50.2% (2.1)
% Difference
Mean (cm)


Positive Fusional Vergence 3m 1-Year Test-Retest
ICC = 0.56
(95% CI: 0.13 to 0.82)
Second Test
First Test
n
1
2
3
4
Positive Fusional Vergence 3m
48.6% (50.8)
Mean = -11.6% (23)
-71.8% (-4.8)
% Difference
Mean (prism diopters)

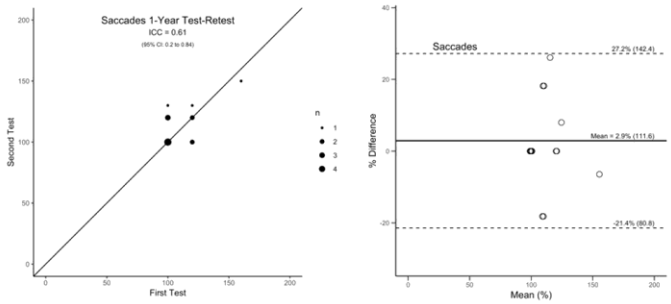

Saccades 1-Year Test-Retest
ICC = 0.61
(95% CI: 0.2 to 0.84)
Second Test
First Test
n
1
2
3
4
Saccades
27.2% (142.4)
Mean = 2.9% (111.6)
-21.4% (80.8)
% Difference
Mean (%)

Figure 4.3. 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 4.2 for explanation of abbreviations and scales.

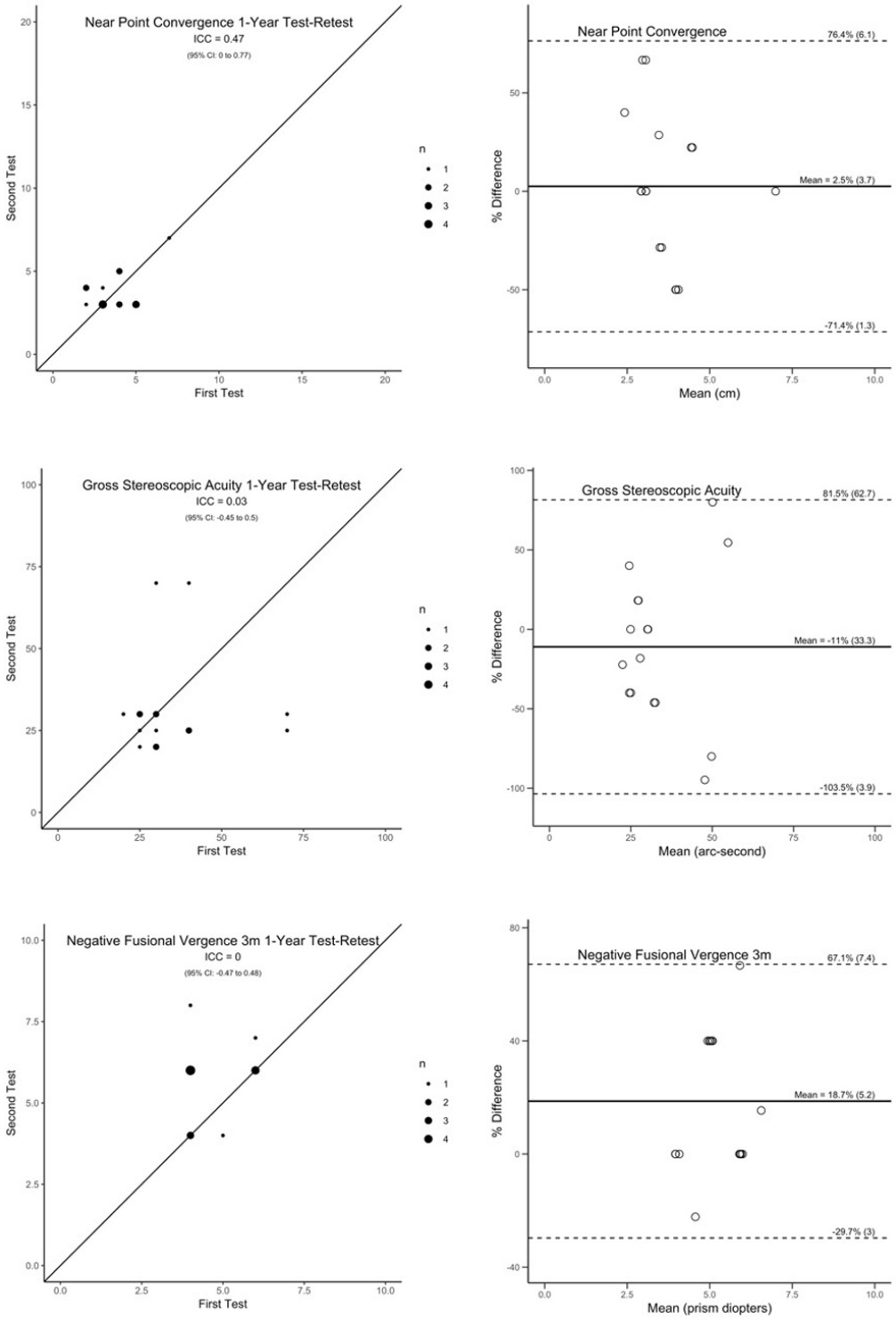


Figure 4.4. Vision tests with poor one-year test-retest reliability.

(**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 4.2 for explanation of abbreviations and scales.

Five tests showed good to moderate one-year test-retest reliability (Figure 4.3), 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 was also one outlier for Positive Fusional Vergence at 3m. When removing this outlier in a sensitivity analysis, the ICC dropped from 0.56 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 values for the measure decreased substantially.

Three of the remaining four tests showed poor one-year test-retest reliability (Figure 4.4). 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 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 [37]. 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]. 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 [18].

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 [9–17, 39-41]. 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 [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 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 4.2) and Positive Fusional Vergence at 3m (Figure 4.3). 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 similar to the value found for the one-week test-retest reliability (ICC=0.54); it did not affect LoA. When we removed the outlier 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 4.3). 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 [38] or esotropia [19] 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 [38]. 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 [16] for data from reference [15]). However, Giffard et al. reported a one-week ICC = 0.84 in patients for NPC with neck pain [39 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 [39] used the RAF rule [24].

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) [40]. 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, 40].

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) [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, 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 [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 (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). 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 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).

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 [36]

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. 27032924 10.3171/2016.1.FOCUS15610

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. 16983222 10.1097/00001199-200609000-00001

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. 21976115

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. 20188017 10.1017/S1355617710000196

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. 23479479 10.1136/bjsports-2013-092313

6 McCrory P, Meeuwisse W, Dvořák J, et al.: Consensus statement on concussion in sport-the 5 th international conference on concussion in sport held in Berlin, October 2016. *Br J Sports Med.* 2017;51(11):838–47. 28446457 10.1136/bjsports-2017-097699

7 Ventura RE, Balcer LJ, Galetta SL: The neuro-ophthalmology of head trauma. *Lancet Neurol.* 2014;13(10):1006–16. 25231523 10.1016/S1474-4422(14)70111-5

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. 20883154 10.1089/neu.2010.1512 3922228

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. 28362161 10.4085/1062-6050-52.2.12 5455247

10 Goss DA, Becker E: Comparison of near fusional vergence ranges with rotary prisms and

with prism bars. *Optometry.* 2011;82(2):104–7. 21144803 10.1016/j.optm.2010.09.011

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. 20199880 10.1016/j.jaapos.2009.12.159 2866770

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. 26453625 10.1177/0363546515606430 5067104

13 Maples WC, Hoenes R: Near point of convergence norms measured in elementary school children. *Optom Vis Sci.* 2007;84(3):224–8. 17435536 10.1097/OPX.0b013e3180339f44

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. 18761485 10.1111/j.1475-1313.2008.00583.x

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. 18905674 10.1016/0002-9394(48)90862-9

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. 11999151 10.1097/00006324-200204000-00012

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. 10.1249/01.mss.0000538190.02670.21

18 Broglio SP, Ferrara MS, Macciocchi SN, et al.: Test-retest reliability of computerized concussion assessment programs. *J Athl Train.* 2007;42(4):509–14. 18174939 2140077

19 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. 19139339 10.1001/archophthalmol.2008.548 2629143

20 Stereo optical co: Randot stereotest. In: Stereo optical co., ed.;1995.

21 Rowe F: Clinical orthoptics. 3rd ed. Chichester, West Sussex: Wiley-Blackwell;2012. 10.1002/9781118702871

22 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

23 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

24 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. 10.1016/B978-0-7506-4713-

7.50007-2

25 Scheiman M, Gwiazda J, Li T: Non-surgical interventions for convergence insufficiency. *Cochrane Database Syst Rev.* 2011; (3):CD006768. 21412896 10.1002/14651858.CD006768.pub2 4278667

26 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. 9703039

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

28 Birch E, Williams C, Drover J, et al.: Randot preschool stereoacuity test: Normative data and validity. *J AAPOS.* 2008;12(1):23–6. 17720573 10.1016/j.jaapos.2007.06.003 2577836

29 Piano ME, Tidbury LP, O'Connor AR: Normative Values for Near and Distance Clinical Tests of Stereoacuity. *Strabismus.* 2016;24(4):169–72. 27929725 10.1080/09273972.2016.1242636

30 Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420–8. 18839484 10.1037/0033-2909.86.2.420

31 Bland JM, Altman D: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1(8476):307–10. 2868172 10.1016/S0140-6736(86)90837-8

32 Koo TK, Li MY: A guideline of selecting and reporting intraclass correlation coefficients for

reliability research. *J Chiropr Med.* 2016;15(2):155–63. 27330520 10.1016/j.jcm.2016.02.012 4913118

33 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. 20937092 10.1186/1477-7525-8-116 2959099

34 Saville DJ: Multiple comparison procedures: The practical solution. *Am Stat.* 1990;44(2):174–80. 10.2307/2684163

35 R core team: R: A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing; 2015.2015. Reference Source

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. 27182767 10.1080/21622965.2016.1138310

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. 25360761 10.3109/09273972.2014.972521

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. 28866427 10.1016/j.msksp.2017.08.010

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. 26675120 10.7860/JCDR/2015/14407.6747 4668442

41 Ettinger U, Kumari V, Crawford TJ, et al.: Reliability of smooth pursuit, fixation, and saccadic eye movements. *Psychophysiology.* 2003;40(4):620–8. 14570169 10.1111/1469-8986.00063

42 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

43 Mainwaring L, Pennock KMF, Mylabathula S et al. Subconcussive head impacts in sport: A systematic review of the evidence. International Journal of Psychophysiology 2018;132: 39-54.

44 Klein C, Fischer B. Instrumental and test-retest reliability of saccadic measures. Biological Psychology 2005;68:201-213.

5 Discussion

5.1 Interpretation of results

A vision test used for the diagnosis and management of concussion must be valid and sensitive to the effect of a concussion (32). Assessing if a test is sensitive to the impact of concussion requires that it can differentiate between a true change in the vision function (signal) or just error (noise). This is assessed with test-retest reliability (32).

In general, test-retest reliability refers to the consistency of tests' measurements (109). In this thesis, I have explored one-year test-retest reliability for the ten vision tests in young elite Canadian athletes, measured by a single clinician at APEXK Inc. APEXK Inc. is a company focused on providing orthoptic examinations for athletes. The one-year test-retest reliability of the tests ranged from poor to excellent, summarized in table 5.1. However, with a more in-depth examination of the data, all the tests had moderate to poor reliability in a one-year interval. Positive Fusional Vergence at 30cm initially showed excellent reliability but, in a subsequent sensitivity analysis, showed only moderate reliability. A moderate to poor level of test-retest reliability may still be useful in the diagnosis and management of a concussion if concussion produces a moderate to a large change in the vision function. In terms of the analogy of telephone communication (section 2.4), even if the noise is substantial in these tests, a strong signal can still be detected if it is larger than the noise. In terms of the agreement, the majority of the tests (except Phoria at 30cm and Saccades) had standardized 95% LoA in the range of 40–90%. This level of the agreement indicates that repeated scores of an individual over time are

expected to vary by 40–90% of the mean score, even if there is no actual change in vision function (signal).

As mentioned in the literature review, no published studies have examined the stability of the ten specific vision tests used in this study over a one-year interval, in athlete or non-athlete populations. Although some studies evaluated the test-retest reliability of Saccades over one-year interval and more (110-112), the data from studies cannot be compared with the results of the tests in this thesis. The tests used in these previous studies were very different from our test of Saccades. These tests assessed very different aspects of Saccades. For instance, some of these studies assessed King-Devick test (110, 111). This is a specialized test that measures the speed of rapid number naming, whereas we performed a general test of Saccades (112).

The time interval for test-retest reliability must fall within a range needed for clinical interpretation of concussion assessment. If athletes have annual pre-season tests, they can have a concussion anytime up to one year after a pre-season test. Therefore, in the context of comparing results after a concussion to annual baseline tests conducted before the concussion, the time-frame for reliability comparisons should be up to one year (32). Although there were no long-term test-retest reliability studies on the ten vision tests used in this study, there are reports of short-term test-retest reliability. I have summarized the results of these studies in table 2.3 of the literature review section of this thesis. These prior studies reported a range of poor to excellent reliability (ICC=0.34 to 1.0) for the tests in a one-week or shorter interval. These studies have employed various methods among various groups of individuals, including healthy and concussed children and adults (18-23, 29, 113). One of these studies by Long et al. has applied

the same measurement techniques on the same ten vision tests we used in this study but were limited to a short reliability window of one-week (25).

I have compared the results of the one-week study by Long et al. with our one-year study in table 5.1, and categorized the differences in ICC and LOA into four theoretical categories based on whether the ICC was lower/equal (expected) vs higher (should not be possible), and whether the LoA was higher/equal (expected) vs. lower (unexpected).

1) The one-year ICC of a test was less than or equal to the one-week interval, and the 95% LoA was higher (Gross Stereoscopic Acuity, NPC). This observation is expected as more variation (more noise) is expected over more extended periods, and any learning effect would be less pronounced over a one-year period.

2) Both the one-year ICC and 95% LoA of a test were less than or equal to the one-week ICC (NPCb, and Negative Fusional Vergence at 3m). Although lower or equal ICC was expected with the one-year interval, we did not expect better agreement in one-year interval compared to one-week interval. One possibility is that the participants of the one-year study were elite athletes while the participants of the one-week study were a healthy population. Elite athletes must be consistent in their performance or they would not achieve elite status. It is possible that high variability in visual function for any reason (e.g. stress, anxiety) would preclude consistent performances in sport. Alternatively, the differences may be unrelated to the population and could have occurred by chance.

3) The one-year ICC of a test was higher than the ICC of the one-week, but the 95% LoA was less than or equal to the one-week (Positive Fusional Vergence at 30cm and 3m, Negative

Fusional Vergence at 30cm and Saccades). Lower 95% LoA was discussed in the previous paragraph. In theory, a higher ICC at one-year compared to one-week is not possible because we expect more variation (more noise) over a more extended period. We identified five following possible reasons for this observation.

First, when we explored the data further, we noticed one outlier that significantly increased the range of values for Positive Fusional Vergence at 30cm (see figure 4.2) and Positive Fusional Vergence at 3m (see figure 4.3). Expanding the range of values increases the ICC by increasing variability between participants. As I acknowledged in the literature review section of this thesis, ICC is based on the results of the analysis of variance (ANOVA). ANOVA separates the error into variability between participants (range of values along x or y axes) and variability within a participant. Therefore, if variability between participants increases by a broader range of values, ICC will increase. We explored how removing the outlier in our data would affect the results. In Positive Fusional Vergence at 30cm, the ICC dropped to 0.53 from 0.93, which is below the value found for the one-week test-retest reliability; however, it did not affect LoA. In addition, when we removed the outlier from Positive Fusional Vergence at 3m (same person as 30cm), the ICC decreased to 0.21 from 0.56. This was again below the one-week ICC. Note that the outlier for this measure had a large difference between the two test scores, and removing such a data point would typically be expected to increase the ICC. The finding that the ICC decreased by removing the outlier indicates how much the range of data can affect the ICC.

Second, the ICC could have been better at one-year if non-athletes in the one-week study had less motivation to perform well on repeated tests. The result of the one-week study was not

essential for the non-athlete participants. In contrast, baseline tests for the elite athletes in the one-year study were conducted as part of their important preseason assessment.

Third, a potential learning effect could occur with retest measurements. However, a learning effect is less likely in our study because the athletes were tested only twice, with a one-year interval. A learning effect is more likely in a one-week interval.

Fourth, our test results were based on clinical charts where the clinician had access to the previous results, which might artificially increase the reliability of the tests. In contrast, the one-week study by Long et al. (25) was a prospective study, and the clinician who performed the test was blinded.

Fifth, the increased ICC in the one-year study could be the impact of chance due to sampling variation.

4) Both the one-year ICC and 95% LoA of a test were higher than the one-week ICC. We did not observe any results where both ICC and LoA were higher.

Table 5-1. Summary of one-week (113) and our one-year test-retest reliability and agreement of the ten vision tests

One-year vs. one-week	Vision tests	One-year ICC	One-year 95% LoA (%)	One-week ICC	One-week 95% LoA (%)
Lower ICC, higher LoA	NPC	0.47	±73.9	0.54	±57.9
	Gross Stereoscopic Acuity	0.03	±92.5	0.86	±55
Lower or similar ICC, lower LoA	Phoria, 30cm	0.68	±119.2	0.69	±122
	NPCb	0.65	±49.4	0.64	±65.2
	Phoria at 3m	0[1]	-	0.88	±123.6
	Negative Fusional Vergence at 3m	0	±48.4	0.43	±68.8
Higher ICC, lower LoA	Positive Fusional Vergence, 30cm	0.93	±41.9	0.54	±69.5
	Negative Fusional Vergence, 30cm	0.78	±41.2	0.66	±63
	Saccades	0.61	±24.3	0.34	±34
	Positive Fusional Vergence, 3m	0.56	±60.2	0.49	±69.8

[1] Most of the measures for Phoria at 3m were identical so both ICC and LoA were not informative.

In the following, I will review individual vision tests.

5.1.1 NPC and NPCb

One-year test-retest reliability of NPC (ICC=0.47) and NPCb (ICC=0.65) were lower than or similar to the results in the one-week reliability study by Long et al. (ICC=0.54 and 0.64, respectively) (25), which was expected as discussed in section 5.1. Other studies have reported higher ICCs (0.65 to 0.97) for NPC in a shorter than one-week interval (23, 83). Similarly, ICCs reported for NPCb in other studies were higher than the ICC we found in a one-year interval. (ICC from 0.84 to 0.92) (24, 26). The discrepancies in the results of the studies could be partially due to differences in testing procedures. For instance, we used the Maples method (29), which is a non-accommodative test, whereas some other studies used an accommodative target, such as RAF rule (26) or Astron International Accommodative Rule (24). NPC and NPCb would be useful over a one-year interval if the concussion leads to a moderate to high change in vision function (moderate to high signal).

5.1.2 Gross Stereoscopic Acuity

The one-year test-retest reliability for Gross Stereoscopic Acuity was poor (ICC=0.03). However, the test had good reliability in non-athlete young adults (ICC=0.86) in the one-week interval (25). Another study employed Titmus stereo fly and Frisby stereo tests in children and revealed excellent reliability for the test (ICC=1.0) (27). This was expected, as discussed in the previous section, 5.1. 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. However, it may still be appropriate for use in a one-week interval.

5.1.3 Phoria

Phoria at 30cm had moderate reliability (ICC=0.68) consistent with the one-week reliability study (ICC=0.69) (25). Other studies in people with strabismus (114) or esotropia (115) have not reported ICC. Therefore, comparing between studies is not possible. In addition, we found that the ICC and LoA were not appropriate measures of reliability for Phoria at 3m because most of our participants had identical values for both measurements. If participants would have a broader range of scores, the ICC might provide meaningful information. Although the ICC is not a good measure, 14/16 athletes had identical values suggesting very little "noise". If these results are confirmed in other studies with larger ranges of values for phoria, even small changes in phoria due to concussion may be identified.

5.1.4 Negative Fusional Vergence

Negative Fusional Vergence had good and poor one-year test-retest reliability, at 30cm (ICC=0.78) and 3m (ICC=0.0). The ICC of the one-week study was smaller than the one-year at 30 cm (0.66), and it was larger at 3m (0.43) (25). This was discussed in section 5.1, in detail. This level of reliability would be useful in conditions where the concussion leads to a moderate to high change in vision function (moderate to high signal).

5.1.5 Positive Fusional Vergence

In the one-year test-retest study, Positive Fusional Vergence had excellent reliability at 30cm (ICC=0.93) and moderate at 3m (ICC=0.56). These values were higher than ICCs in one-week test-retest, 0.54 and 0.49, respectively (25). The ICCs were also higher than the ICCs in other one-week studies at 30cm (ICC=0.53 and 0.59), when two examiners assessed participants (24), and distance (ICC=0.72) (23). Our finding was unexpected, as discussed in the previous

section. With sensitivity analysis, estimated ICCs for Positive Fusional Vergence at 30cm and 3m dropped to 0.53 and 0.21, respectively. These ICC were equal or below the ICCs found for the test at the one-week studies (23-25). This was what we expected. With a poor to moderate reliability, Positive Fusional Vergence may be useful in a one-year interval if the concussion causes a significant change in the test score (high signal).

5.1.6 Saccades

The APEXK, the test of Saccades, had moderate reliability (ICC=0.61) with the lowest LoA (in percentage) of the other tests. These findings were better than one-week study findings (25) and similar to findings in healthy adults over about a two-month interval (ICC=0.59) (28). This unexpected finding also has been discussed in the previous section in detail. With moderate reliability and an LoA of ±34%, the results of the test of Saccades could be considered stable over one year. However, as mentioned in the literature review, the test of Saccades in this study was very different from other reports. Therefore, a comparison between our findings and finding from other studies is not possible.

5.2 Summary

In this thesis, five vision tests (NPCb, Negative Fusional Vergence at 30cm, Phoria at 30cm, Positive Fusional Vergence at 30cm, and test of Saccades) had moderate one-year test-retest reliability. This level of reliability would be useful in conditions where the concussion leads to a moderate change in vision function (moderate signal). In addition, four vision tests (Gross Stereoscopic Acuity, NPC, Positive Fusional Vergence at 3m, and Negative Fusional Vergence at 3m) had poor one-year test-retest reliability. These tests may be useful over a one-

year period if the effect of concussion would be large on the vision function (significant signal). The last vision test with identical scores in 14/16 athletes was Phoria at 3m. If test-retest results are the same over a larger range of values of phoria in future studies, even small effects of concussion would be detectable.

5.3 Strengths and limitations

This study has multiple strengths. First, several studies have previously evaluated the inter-rater reliability of some vision tests (114, 115). Still, inter-rater reliability is less important in the context of clinical care when one clinician follows patients over time. The study of this thesis evaluated the test-retest reliability of the ten vision tests over an interval that allows the normal variation over time expected in clinical practice between baseline measures and possible subsequent concussions. Second, we used two statistical analysis methods to assess the test-retest reliability of these vision tests over one-year. The ICC represents how much 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 measure. Furthermore, the 95% LoA for each test provides the magnitude of the normal variation that can be expected with repeated measures. Differences in the 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). Finally, we conducted a 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.

The study has the following limitations. It is a historical cohort observational study. In this method of research, a selected group of participants are followed back in time (116). In these studies, researchers sample a source population to determine their study population (a subset of participants eligible for the study) (117). The disadvantage of historical cohort studies is that the investigator has limited control over data collection. The existing data may be incomplete, inaccurate, or inconsistent. Therefore, historical cohort studies are susceptible to recall, information and selection bias (116, 118).

We have tried to minimize the chance of information and recall bias in our study. We used two sources of data in this study. First, we used electronic medical charts of the chief medical officer of the Institut National du Sport du Quebec (INSQ) to obtain demographic data on athletes with concussion and information on the diagnosis of concussions. Second, we collected vision data for athletes with concussions from the electronic files of a clinician trained in orthoptics at APEXK Inc. Then, we compared all data from the two sources. If we found a discrepancy in the data, we followed a process to clarify this discrepancy that is described in the method section. In the end, both the sports medicine physician and the clinician in APEXK approved the corrected data. In addition, the test results were based on clinical charts where the clinician had access to the previous results, which might artificially increase the reliability of the tests.

There was a chance of recall bias, especially when athletes were asked their age, any history of concussions, etc. in the APEXK data. To minimize the chance of recall bias, we checked the most reliable resource for that specific data and rechecked the data with both INSQ and APEXK clinicians when required. For instance, to identify the age of athletes, we looked at the INSQ data, which had athletes' age based on medicare information. In addition, to reduce the

chance of recall bias about the history of concussion, we compared the APEXK data to INSQ data, where the sports medicine physician diagnosed a concussion.

Another disadvantage of cohort studies is a chance of selection bias, which is less likely in this study because we included all athletes who met the inclusion criteria of two baseline tests with no intervening concussion. We started with all INSQ athletes who had vision test results at APEXK. The vision data was available for 199 athletes. Among these, 181 athletes were excluded because they did not have baseline tests, and two athletes were excluded because their baseline tests were out of the required period of study. Many athletes never went for testing, and some athletes only had one baseline test. Therefore, we had a relatively small sample size of 16 healthy athletes. 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 discover ICC values >=0.7 (Table 1a in (119)). 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 (119)). Finally, the results of the test of Saccades in this study were based on the unpublished proprietary algorithm developed by the clinician at APEXK. This limits its applicability.

5.4 Areas for future research

The current study has defined the test-retest reliability of ten vision tests in young athletes. The next step to assess the utility of these vision tests in the diagnosis and management of concussion is investigating the effect of concussion on the test scores (signal). If a test score changes more than the 95%LoA after a concussion compared to baseline, it may be useful clinically. In addition, any “noise” associated with the test is expected to cause results to vary in

both directions compared to baseline. Therefore, a consistently higher (or lower) score with repeated tests over weeks (e.g. when following a patient with a concussion over time) compared to baseline would suggest noise is not the reason for the change in the value of the test. A further step in the research should look at the results of the tests after recovery from a concussion. These findings will help clinicians and researchers determine the utility of the vision tests in the assessment and follow-up of a concussion.

5.5 Conclusions

In this research, we found five out of the ten vision tests, including Negative Fusional Vergence at 30cm, Phoria at 30cm, NPCb, Positive Fusional Vergence at 30m, and Saccades had good to moderate one-year test-retest reliability. This level of reliability is useful in conditions that produce a moderate change in visual function. The four vision tests, including Negative Fusional Vergence at 3m, NPC, Positive Fusional Vergence at 30cm and Gross Stereoscopic Acuity, had poor one-year test-retest reliability. They may be useful in detecting large effects on visual function. If further studies suggest that the impact of concussion on test scores is moderate to large, these vision tests may still be useful clinically. The limited range of values for Phoria at 3m in our study limits our ability to make general conclusions for this test.

6 References

1. Sussman ES, Ho AL, Pendharkar AV, Ghajar J. Clinical evaluation of concussion: the evolving role of oculomotor assessments. Neurosurg Focus. 2016;40(4):E7.

2. Nygren-de Boussard C, Holm LW, Cancelliere C, Godbolt AK, Boyle E, Stålnacke B-M, et al. Nonsurgical interventions after mild traumatic brain injury: a systematic review. Results of the International Collaboration on Mild Traumatic Brain Injury Prognosis. Archives of physical medicine and rehabilitation. 2014;95(3):S257-S64.

3. Gordon KE, Dooley JM, Wood EP. Descriptive epidemiology of concussion. Pediatric neurology. 2006;34(5):376-8.

4. Canada Go. Concussion in Sport 2018 [Available from: https://www.canada.ca/en/public-health/services/diseases/concussion-sign-symptoms/concussion-sport-infographic.html.

5. Mannix R, O'Brien MJ, Meehan WP, 3rd. The epidemiology of outpatient visits for minor head injury: 2005 to 2009. Neurosurgery. 2013;73(1):129-34; discussion 34.

6. McCrea M, Hammeke T, Olsen G, Leo P, Guskiewicz K. Unreported concussion in high school football players: implications for prevention. Clinical journal of sport medicine. 2004;14(1):13-7.

7. Torres DM, Galetta KM, Phillips HW, Dziemianowicz EM, Wilson JA, Dorman ES, et al. Sports-related concussion: Anonymous survey of a collegiate cohort. Neurol Clin Pract. 2013;3(4):279-87.

8. McCrory P, Meeuwisse WH, Aubry M, Cantu B, Dvořák J, Echemendia RJ, 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.

9. 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.

10. McCrory P, Meeuwisse W, Dvorak J, Aubry M, Bailes J, Broglio S, et al. Consensus statement on concussion in sport-the 5(th) international conference on concussion in sport held in Berlin, October 2016. Br J Sports Med. 2017;51(11):838-47.

11. Whitney SL, Eagle SR, Marchetti G, Mucha A, Collins MW, Kontos AP. Association of acute vestibular/ocular motor screening scores to prolonged recovery in collegiate athletes following sport-related concussion. Brain Injury. 2020;34(6):840-5.

12. Dikmen S, Machamer J, Fann JR, Temkin NR. Rates of symptom reporting following traumatic brain injury. J Int Neuropsychol Soc. 2010;16(3):401-11.

13. Damji F, Babul S. Improving and standardizing concussion education and care: a Canadian experience. Concussion (London, England). 2018;3(4):Cnc58.

14. Unit BIRaP. Concussion [Available from: https://www.injuryresearch.bc.ca/quick-facts/concussion/.

15. Ventura RE, Balcer LJ, Galetta SL. The neuro-ophthalmology of head trauma. Lancet Neurol. 2014;13(10):1006-16.

16. Ventura RE, Balcer LJ, Galetta SL. The Concussion Toolbox: The Role of Vision in the Assessment of Concussion. Semin Neurol. 2015;35(5):599-606.

17. Ventura RE, Jancuska JM, Balcer LJ, Galetta SL. Diagnostic tests for concussion: is vision part of the puzzle? J Neuroophthalmol. 2015;35(1):73-81.

18. Goss DA, Becker E. Comparison of near fusional vergence ranges with rotary prisms and with prism bars. Optometry. 2011;82(2):104-7.

19. Wang J, Hatt SR, O'Connor AR, Drover JR, Adams R, Birch EE, et al. Final version of the Distance Randot Stereotest: normative data, reliability, and validity. J AAPOS. 2010;14(2):142-6.

20. 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.

21. Pearce KL, Sufrinko A, Lau BC, Henry L, Collins MW, Kontos AP. 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.

22. Antona B, Barrio A, Barra F, Gonzalez E, Sanchez I. Repeatability and agreement in the measurement of horizontal fusional vergences. Ophthalmic Physiol Opt. 2008;28(5):475-91.

23. Brozek J, Simonson E, et al. Effects of practice and the consistency of repeated measurements of accommodation and vergence. Am J Ophthalmol. 1948;31(2):191-8.

24. Rouse M, Borsting E, Deland P, The Convergence Insufficiency and Reading Study (CIRS) Group. Reliability of binocular vision measurements used in the classification of convergence insufficiency. Optom Vis Sci. 2002;79(4):254-64.

25. Long S, Leclerc S, Tinjust D, Steele R, Schuster T, Shrier I. Determining Consistency And Agreement Of Scores Across Two Measurements Of The Visual System: Test-retest Reliability: 2730 Board #13 June 1 2: 00 PM - 3: 30 PM. Medicine & Science in Sports & Exercise. 2018;50(5S):664.

26. Giffard P, Daly L, Treleaven J. Influence of neck torsion on near point convergence in subjects with idiopathic neck pain. Musculoskeletal science & practice. 2017;32:51-6.

27. Moganeswari D, Thomas J, Srinivasan K, Jacob GP. Test Re-Test Reliability and Validity of Different Visual Acuity and Stereoacuity Charts Used in Preschool Children. Journal of clinical and diagnostic research : JCDR. 2015;9(11):Nc01-5.

28. Ettinger U, Kumari V, Crawford TJ, Davis RE, Sharma T, Corr PJ. Reliability of smooth pursuit, fixation, and saccadic eye movements. Psychophysiology. 2003;40(4):620-8.

29. Maples WC, Hoenes R. Near point of convergence norms measured in elementary school children. Optom Vis Sci. 2007;84(3):224-8.

30. Guskiewicz KM, Bruce SL, Cantu RC, Ferrara MS, Kelly JP, McCrea M, et al. National Athletic Trainers' Association Position Statement: Management of Sport-Related Concussion. J Athl Train. 2004;39(3):280-97.

31. Graham R, Rivara FP, Ford MA, Spicer CM, Youth CoS-RCi, Council NR. Concussion recognition, diagnosis, and acute management. Sports-Related Concussions in Youth: Improving the Science, Changing the Culture: National Academies Press (US); 2014.

32. Broglio SP, Ferrara MS, Macciocchi SN, Baumgartner TA, Elliott R. Test-retest reliability of computerized concussion assessment programs. J Athl Train. 2007;42(4):509-14.

33. Whiteneck GG, Cuthbert JP, Corrigan JD, Bogner JA. Risk of negative outcomes after traumatic brain injury: a statewide population-based survey. The Journal of head trauma rehabilitation. 2016;31(1):E43-E54.

34. Peterson AB, Xu L, Daugherty J, Breiding MJ. Surveillance report of traumatic brain injury-related emergency department visits, hospitalizations, and deaths, United States, 2014. 2019.

35. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. The Lancet. 1974;304(7872):81-4.

36. Maas AI, Stocchetti N, Bullock R. Moderate and severe traumatic brain injury in adults. The Lancet Neurology. 2008;7(8):728-41.

37. Kay A, Teasdale G. Head injury in the United Kingdom. World journal of surgery. 2001;25(9):1210-20.

38. Kay T, Harrington DE, Adams R, Anderson T, Berrol S, Cicerone K, et al. Definition of mild traumatic brain injury. Journal of Head Trauma Rehabilitation. 1993;8(3):86-7.

39. McKinlay A, Grace R, Horwood L, Fergusson D, Ridder EM, MacFarlane M. Prevalence of traumatic brain injury among children, adolescents and young adults: prospective evidence from a birth cohort. Brain injury. 2008;22(2):175-81.

40. Taylor AM, Nigrovic LE, Saillant ML, Trudell EK, Proctor MR, Modest JR, et al. Trends in ambulatory care for children with concussion and minor head injury from eastern Massachusetts between 2007 and 2013. The Journal of pediatrics. 2015;167(3):738-44.

41. Arbogast KB, Curry AE, Pfeiffer MR, Zonfrillo MR, Haarbauer-Krupa J, Breiding MJ, et al. Point of health care entry for youth with concussion within a large pediatric care network. JAMA pediatrics. 2016;170(7):e160294-e.

42. Rao Deepa P, Steven M, Wendy T. At-a-glance-traumatic brain injury management in Canada: changing patterns of care. Health promotion and chronic disease prevention in Canada: research, policy and practice. 2018;38(3):147.

43. Lumba-Brown A, Yeates KO, Sarmiento K, Breiding MJ, Haegerich TM, Gioia GA, et al. Centers for Disease Control and Prevention Guideline on the Diagnosis and Management of Mild Traumatic Brain Injury Among Children. JAMA Pediatr. 2018;172(11):e182853.

44. Harmon KG, Drezner JA, Gammons M, Guskiewicz KM, Halstead M, Herring SA, et al. American Medical Society for Sports Medicine position statement: concussion in sport. Br J Sports Med. 2013;47(1):15-26.

45. STATEMENTS Q. VA/DoD clinical practice guideline for management of concussion/mild traumatic brain injury. 2009.

46. McCrea M, Guskiewicz KM, Marshall SW, Barr W, Randolph C, Cantu RC, et al. Acute effects and recovery time following concussion in collegiate football players: the NCAA Concussion Study. Jama. 2003;290(19):2556-63.

47. McCrea M, Guskiewicz K, Randolph C, Barr WB, Hammeke TA, Marshall SW, et al. Incidence, clinical course, and predictors of prolonged recovery time following sport-related concussion in high school and college athletes. Journal of the International Neuropsychological Society. 2013;19(1):22-33.

48. Collins M, Lovell MR, Iverson GL, Ide T, Maroon J. Examining concussion rates and return to play in high school football players wearing newer helmet technology: a three-year prospective cohort study. Neurosurgery. 2006;58(2):275-86; discussion -86.

49. Leddy JJ, Haider MN, Hinds AL, Darling S, Willer BS. A Preliminary Study of the Effect of Early Aerobic Exercise Treatment for Sport-Related Concussion in Males. Clinical Journal of Sport Medicine. 2019;29(5):353-60.

50. Schneider KJ, Leddy JJ, Guskiewicz KM, Seifert T, McCrea M, Silverberg ND, et al. Rest and treatment/rehabilitation following sport-related concussion: a systematic review. Br J Sports Med. 2017;51(12):930-4.

51. Lumba-Brown A, Yeates KO, Sarmiento K, Breiding MJ, Haegerich TM, Gioia GA, et al. Centers for Disease Control and Prevention guideline on the diagnosis and management of mild traumatic brain injury among children. JAMA pediatrics. 2018;172(11):e182853-e.

52. Gravel J, D'Angelo A, Carrière B, Crevier L, Beauchamp MH, Chauny J-M, et al. Interventions provided in the acute phase for mild traumatic brain injury: a systematic review. Systematic reviews. 2013;2(1):63.

53. Zemek RL, Farion KJ, Sampson M, McGahern C. Prognosticators of persistent symptoms following pediatric concussion: a systematic review. JAMA pediatrics. 2013;167(3):259-65.

54. DuPrey KM, Webner D, Lyons A, Kucuk CH, Ellis JT, Cronholm PF. Convergence insufficiency identifies athletes at risk of prolonged recovery from sport-related concussion. The American journal of sports medicine. 2017;45(10):2388-93.

55. Association AP, Association AP. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC: American psychiatric association; 2013.

56. Iverson GL, Gardner AJ, Terry DP, Ponsford JL, Sills AK, Broshek DK, et al. Predictors of clinical recovery from concussion: a systematic review. Br J Sports Med. 2017;51(12):941-8.

57. Cantu RC. Second-impact syndrome. Clinics in sports medicine. 1998;17(1):37-44.

58. McCrory PR, Berkovic SF. Second impact syndrome. Neurology. 1998;50(3):677-83.

59. Guskiewicz KM, McCrea M, Marshall SW, Cantu RC, Randolph C, Barr W, et al. Cumulative effects associated with recurrent concussion in collegiate football players: the NCAA Concussion Study. Jama. 2003;290(19):2549-55.

60. Shrier I, Piché A, Steele RJ. First concussion did not increase the risk of subsequent concussion when patients were managed appropriately. BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine; 2019.

61. Sigurdardottir S, Andelic N, Roe C, Jerstad T, Schanke A-K. Post-concussion symptoms after traumatic brain injury at 3 and 12 months post-injury: A prospective study. Brain Injury. 2009;23(6):489-97.

62. Feddermann-Demont N, Echemendia RJ, Schneider KJ, Solomon GS, Hayden KA, Turner M, et al. What domains of clinical function should be assessed after sport-related concussion? A systematic review. Br J Sports Med. 2017;51(11):903-18.

63. Echemendia RJ, Meeuwisse W, McCrory P, Davis GA, Putukian M, Leddy J, et al. The sport concussion assessment tool 5th edition (SCAT5): background and rationale. British Journal of Sports Medicine. 2017;51(11):848-50.

64. Gunasekaran P, Hodge C, Rose K, Fraser CL. Persistent visual disturbances after concussion. Australian journal of general practice. 2019;48(8):531.

65. Mucha A, Collins MW, Elbin RJ, Furman JM, Troutman-Enseki C, DeWolf RM, et al. A Brief Vestibular/Ocular Motor Screening (VOMS) assessment to evaluate concussions: preliminary findings. Am J Sports Med. 2014;42(10):2479-86.

66. Master CL, Scheiman M, Gallaway M, Goodman A, Robinson RL, Master SR, et al. Vision diagnoses are common after concussion in adolescents. Clinical pediatrics. 2016;55(3):260-7.

67. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. Journal of clinical epidemiology. 2006;59(10):1033-9.

68. Hernaez R. Reliability and agreement studies: a guide for clinical investigators. Gut. 2015;64(7):1018-27.

69. Kottner J, Streiner DL. The difference between reliability and agreement. Journal of clinical epidemiology. 2011;64(6):701-2.

70. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? Physiotherapy. 2000;86(2):94-9.

71. DeVellis RF. Classical test theory. Medical care. 2006:S50-S9.

72. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological methods. 1996;1(1):30.

73. Zaki R, Bulgiba A, Nordin N, Azina Ismail N. A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables. Iran J Basic Med Sci. 2013;16(6):803-7.

74. STREINER D. Health measurement scales. A practical guide to their development and use. 1995:104–27.

75. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin. 1979;86(2):420.

76. Trevethan R. Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. Health Services and Outcomes Research Methodology. 2017;17(2):127-43.

77. Lord FM, Novick MR. Statistical theories of mental test scores 1968 Reading. MA Addison-Wesley.

78. Giavarina D. Understanding Bland Altman analysis. Biochemia medica. 2015;25(2):141-51.

79. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;8(2):135-60.

80. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Journal of the Royal Statistical Society: Series D (The Statistician). 1983;32(3):307-17.

81. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet. 1986;327(8476):307-10.

82. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet (London, England). 1995;346(8982):1085-7.

83. Pearce KL, Sufrinko A, Lau BC, Henry L, Collins MW, Kontos AP. Near point of convergence after a sport-related concussion: measurement reliability and relationship to neurocognitive impairment and symptoms. The American journal of sports medicine. 2015;43(12):3055-61.

84. Wardle SG, Cass J, Brooks KR, Alais D. Breaking camouflage: binocular disparity reduces contrast masking in natural images. J Vis. 2010;10(14).

85. Wardle SG, Bex PJ, Cass J, Alais D. Stereoacuity in the periphery is limited by internal noise. J Vis. 2012;12(6):12.

86. Birch E, Williams C, Drover J, Fu V, Cheng C, Northstone K, et al. Randot Preschool Stereoacuity Test: normative data and validity. J AAPOS. 2008;12(1):23-6.

87. Piano ME, Tidbury LP, O'Connor AR. Normative Values for Near and Distance Clinical Tests of Stereoacuity. Strabismus. 2016;24(4):169-72.

88. Millodot M. Dictionary of optometry and visual science (5: e uppl.). s 107. Elsevier Butterworth-Heinemann; 2000.

89. Von Noorden G. Examination of patient-II: Binocular vision and ocular motility, 187-189, Mosby, St. Louis; 1996.

90. Siderov J, Chiu SC, Waugh SJ. Differences in the nearpoint of convergence with target type. Ophthalmic Physiol Opt. 2001;21(5):356-60.

91. 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.

92. Scheiman M, Gallaway M, Frantz KA, Peters RJ, Hatch S, Cuff M, et al. Nearpoint of convergence: test procedure, target selection, and normative data. Optometry and Vision Science. 2003;80(3):214-25.

93. ROUSE MW, BORSTING E, DELAND PN, Insufficiency aTC, Group RS. Reliability of Binocular Vision Measurements Used in the Classification of Convergence Insufficiency. Optometry and Vision Science. 2002;79(4):254-64.

94. Wilmer JB, Backus BT. Genetic and environmental contributions to strabismus and phoria: Evidence from twins. Vision Research. 2009;49(20):2485-93.

95. Calkins ME, Iacono WG. Eye movement dysfunction in schizophrenia: a heritable characteristic for enhancing phenotype definition. American Journal of Medical Genetics. 2000;97(1):72-6.

96. Crevits L, Hanse M, Tummers P, Van Maele G. Antisaccades and remembered saccades in mild traumatic brain injury. Journal of neurology. 2000;247(3):179-82.

97. Heitger MH, Jones RD, Macleod AD, Snell DL, Frampton CM, Anderson TJ. Impaired eye movements in post-concussion syndrome indicate suboptimal brain function beyond the influence of depression, malingering or intellectual ability. Brain. 2009;132(Pt 10):2850-70.

98. 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.

99. Scheiman M, Gwiazda J, Li T. Non-surgical interventions for convergence insufficiency. Cochrane Database Syst Rev. 2011(3):CD006768.

100. Hayes GJ, Cohen BE, Rouse MW, De Land PN. Normative values for the nearpoint of convergence of elementary schoolchildren. Optom Vis Sci. 1998;75(7):506-12.

101. Sutter P, Harvey L. Vision Rehabilitation: Multidisciplinary Care of the Patient Following Brain Injury. Boca Raton: Taylor & Francis Group; 2011.

102. Rowe F. Clinical Orthoptics. 3rd ed. Chichester, West Sussex: Wiley-Blackwell; 2012.

103. 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.

104. . !!! INVALID CITATION !!! .

105. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine. 2016;15(2):155-63.

106. Johnston BC, Thorlund K, Schünemann HJ, Xie F, Murad MH, Montori VM, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. Health and quality of life outcomes. 2010;8:116.

107. Saville DJ. Multiple comparison procedures: the practical solution. The American Statistician. 1990;44(2):174-80.

108. Core Team R. R: A Language and Environment for Statistical Computing; 2015. R Foundation for Statistical Computing: Vienna. 2015.

109. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. Journal of strength and conditioning research. 2005;19 1:231-40.

110. Breedlove KM, Ortega JD, Kaminski TW, Harmon KG, Schmidt JD, Kontos AP, 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. Journal of athletic training. 2019;54(12):1241-6.

111. 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):2283-6.

112. Klein C, Fischer B. Instrumental and test-retest reliability of saccadic measures. Biological psychology. 2005;68(3):201-13.

113. Rouse MW, Borsting E, Deland PN. Reliability of binocular vision measurements used in the classification of convergence insufficiency. Optom Vis Sci. 2002;79(4):254-64.

114. de Jongh E, Leach C, Tjon-Fo-Sang MJ, Bjerre A. 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.

115. Interobserver reliability of the prism and alternate cover test in children with esotropia. Archives of ophthalmology (Chicago, Ill : 1960). 2009;127(1):59-65.

116. Elwood M. Critical appraisal of epidemiological studies and clinical trials: Oxford University Press; 2017.

117. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plastic and reconstructive surgery. 2010;126(6):2234-42.

118. Hulley SB. Designing clinical research: Lippincott Williams & Wilkins; 2007.

119. Bujang MA, Baharum N. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. Archives of Orofacial Science. 2017;12(1).