

Running title: Concussion

1 *One-week test-retest reliability of nine binocular tests and saccades used*
2 *in concussion*

3

4 Stephanie Long MSc¹, Tibor Schuster PhD¹, Russell Steele PhD², Suzanne Leclerc MD PhD³,
5 Ian Shrier MD PhD^{1,4}

6 Author affiliations: ¹Department of Family Medicine, McGill University; ²Department of
7 Mathematics and Statistics, McGill University; ³Institut National du Sport du Quebec, Montreal,
8 Canada; ⁴Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital,
9 McGill University, Montreal, Canada

10

11 Corresponding Author:

12 Ian Shrier MD, PhD

13 Centre for Clinical Epidemiology,

14 Lady Davis Institute, Jewish General Hospital

15 3755 Cote Sainte-Catherine Road

16 Montreal, QC H3T 1E2

17 Canada

18 Email: ian.shrier@mcgill.ca

19

Running title: Concussion

20 **Abstract**

21 **Background:** Tests of binocular vision (BVTs) and ocular motility are used in concussion
22 assessment and management.

23 **Purpose:** To determine the one-week test-retest reliability of 9 binocular vision tests (BVTs) and
24 a test of saccades proposed for use in concussion management.

25 **Study Design:** Prospective test-retest.

26 **Methods:** We examined the one-week test-retest reliability of 9 BVTs in healthy participants: 3D
27 vision (gross stereoscopic acuity), phoria at 30cm and 3m, ability of eyes to move/fixate in-sync
28 (positive and negative fusional vergence at 30cm and 3m, near point of convergence and near
29 point of convergence – break [i.e. double vision]) and 1 ocular motor test, saccades.

30 **Results:** We tested 10 males and 10 females without concussion and a mean age of 25.5 (4.1)
31 years. The intraclass correlations suggest good reliability for phoria 3m (0.88) and gross
32 stereoscopic acuity (0.86), and moderate reliability for phoria 30cm (0.69), near point of
33 convergence (0.54), positive fusional vergence (0.54) and negative fusional vergence (0.66) at
34 30cm, and near point of convergence - break (0.64). There was poor reliability for saccades
35 (0.34), and both positive and negative fusional vergence (0.49 and 0.43, respectively) at 3m.
36 Limits of agreement (LoA) were best for saccade ($\pm 34\%$) and worst for phoria 30 cm ($\pm 121\%$)
37 and ranged from $\pm 58\%$ to $\pm 70\%$ for 7 of the 8 other tests. The LoA for phoria at 3m were
38 uninformative because measurements for 18 of 20 participants were identical.

39 **Conclusion:** We found test-retest reliability of the BVTs and saccades ranging from poor to
40 good in healthy participants, with the majority being moderate.

41 **Clinical Relevance:** For these vision tests to be clinically useful, the effect of concussion must
42 have a moderate to large effect on the scores of most of the tests.

Running title: Concussion

43 *Key Words:*

44 concussion, vision, near point convergence, reliability

45

46 *What is known about the subject:*

- 47
- Concussions may affect some parts of visual function
- 48
- 1-week test-retest reliability for most visual tests is under-studied

49 *What this study adds to existing knowledge:*

- 50
- We provide intra-class coefficients and limits of agreement for 10 different visual function
- 51
- tests commonly conducted by clinicians in patients with concussion.

52

53

Running title: Concussion

54 *Introduction*

55 Many brain-related disorders (e.g. concussion,²⁹ Parkinson's Disease,⁶ attention deficit
56 hyperactivity disorder,²⁰ stroke⁴⁴) have a visual component as part of their findings.²⁴ For
57 example, posttraumatic vision impairments have been reported in 30% to 65% of patients with a
58 mild traumatic brain injury,¹¹ and are found in nearly 30% of patients with a sport-related
59 concussion.²⁵ Some symptoms associated with concussion are believed to be caused by
60 deficits in the visual system and include: headaches, sensitivity to light, diplopia, and blurred
61 vision.¹²

62 Tests of binocular vision (BVTs) and ocular motility over the last 70 years⁹ include^{9,23,37}: gross
63 stereoscopic acuity, near point of convergence - break (i.e. double vision), phoria at 30cm and
64 3m, positive and negative fusional vergence at 30cm and 3m, and saccadic eye movement
65 assessment. Recent studies have suggested that concussion may result in deficits of
66 convergence, binocular vision, and ocular motility.^{4,12,45} Convergence insufficiency is a disorder
67 of binocular vision diagnosed by abnormal near point of convergence, one of many BVTs that
68 assesses an individual's visual capacity.

69 Before we can conclude that visual function is abnormal in concussion, we must first understand
70 the test-retest reliability of vision tests in healthy control participants. Despite their frequent use,
71 limited studies have examined the reliability of these vision tests. When evaluating the one-
72 week test-retest reliability of the Randot Stereotest (test of gross stereoscopic acuity), 71-82%
73 of adult and child participants with normal vision and with strabismus had perfect agreement
74 between two measurements;^{17,48} common psychometric properties such as intraclass
75 correlation coefficient (ICC) or limits of agreement (LoA) were not reported. An older study of
76 near point of convergence reported an ICC of 0.65 across six testing sessions in six healthy
77 adults, but failed to specify the number of examiners or the time interval between testing

Running title: Concussion

78 sessions.¹⁰ A study of near point of convergence – break in school-aged children reported
79 excellent reliability ($ICC \geq 0.94$).³⁶ A more recent study of three consecutive measurements of
80 near point of convergence – break without a rest interval between tests, found ICCs ranging
81 from 0.78-0.89 in young concussed athletes with convergence insufficiency and 0.92-0.97 in
82 concussed athletes with normal vision.³¹ When examining phoria with the prism alternate cover
83 test, one study reported 95% LoA of ± 4.1 to ± 7.3 prism diopters for distance and ± 3.3 to ± 8.3
84 prism diopters for near in young children with esotropia, but did not provide ICC.³² A one-week
85 test-retest reliability study assessing positive fusional vergence with a prism bar reported ICCs
86 ranging from 0.53-0.59.³⁶ One to ten day test-retest reliability of the prism bar test had 95% LoA
87 ± 4.0 for negative fusional vergence and ± 13.9 for positive fusional vergence.⁵ Finally, the only
88 study we could find that measured reliability for saccades used a computerized prosaccade
89 task. The authors reported moderate two-month test-retest reliability ($ICC = 0.59$) in adult
90 participants with normal vision.¹⁶

91 The above studies provide some information regarding the reliability of these vision tests.
92 However, vision tests are used to follow patients over time, and one might expect additional
93 variability when patients are measured on different days due to fatigue, stress, and other
94 factors. Understanding the usual variability that is independent of changes in pathology or
95 recovery is essential for proper interpretation of trends in results over time. Therefore, the
96 objective of this study was to determine the one-week test-retest reliability of 9 BVTs and a test
97 of saccades in healthy adult participants. We evaluated versions of these vision tests that are
98 commonly used by clinicians as opposed to versions used in research studies in order to best
99 assess the utility of the tests in clinical practice. All tests in the present study were performed by
100 the same clinician on two testing dates.

Running title: Concussion

101 *Methods*

102 *Study Design*

103 We examined a convenience sample of healthy colleagues, friends, and social connections in
104 Montreal, Canada on two separate occasions exactly seven days apart at approximately the
105 same time of day (e.g. morning vs. evening). One individual clinician trained in orthoptics
106 examined all participants individually. We arranged for 4-6 participants to be examined
107 sequentially in a two to three-hour block of time. We randomized the order in which participants
108 were examined using a random number generator within each block. To minimize the probability
109 of the clinician recalling the first scores at the second visit: (1) the clinician verbally reported the
110 results to a research assistant (recorded on paper) and was not informed of any scores until
111 data collection was completed, and (2) at the second visit, the same group of 4-6 participants
112 was examined over the same block of time but their examination order was changed compared
113 to the first visit. This study was approved by the Jewish General Hospital Institutional Review
114 Board.

115 *Participant Selection*

116 We included healthy adults 18 to 35 years. There were no participants with a history of
117 conditions that may affect BVTs (or treatment for such conditions) or saccades such as
118 strabismus (contraindication to BVTs), migraines, neurological disorders, or currently taking
119 muscle relaxants, selective serotonin-reuptake inhibitors, anxiolytics, stimulants, or any other
120 drug class for other psychological conditions that might affect test results. Although there were
121 three subjects with remote history of concussion, they had fully recovered and were not
122 experiencing concussion symptoms or related limitations at the time of our study.

Running title: Concussion

123 *Clinical procedures and measures*

124 We collected the following demographic information at the first visit: date of birth, sex, highest
125 level of education achieved, use of corrective lenses, occupation, and any relevant past medical
126 history (e.g. migraines, vision problems, use of medication, history of concussion).

127 Prior to conducting the vision tests, each participant first completed the symptom portion of the
128 validated sport concussion assessment tool (SCAT3) form.^{1,51} These results were used as a
129 sensitivity analysis to evaluate if changes in their physical states at the two testing sessions
130 might explain potential discrepancies.

131 *Vision Tests*

132 We examined 9 BVTs and a test of saccades. For gross stereoscopic acuity, near point of
133 convergence, near point of convergence – break and phoria, a lower score represents better
134 vision function. For positive and negative fusional vergence, and saccades, a higher score
135 represents better vision function. Brief descriptions of each are provided below and more details
136 are available in the *Appendix*.

137 For gross stereoscopic acuity, we used the Randot Stereotest (Stereo Optical Co., Inc.,
138 Chicago, IL) according to manufacturer's instructions.⁴³ For all tests using a target, the tip of a
139 ballpoint pen was used as the near target, and a 6cm² square card mounted on a wall was used
140 as the far target. For near point of convergence and near point of convergence - break, we
141 followed procedures by Maples et al.²⁷ The near point of convergence score was the distance
142 (cm) between the bridge of the nose and the target at the closest point at which the individual
143 could maintain balanced oculomotor synergy between both eyes, which is identified as when
144 one eye diverges outwards.⁹ The near point of convergence – break score was the distance
145 between the bridge of the nose and the point at which diplopia occurred.^{7,23} We measured
146 phoria at 30cm and 3m using the prism alternate cover test with procedures described by the

Running title: Concussion

147 Pediatric Eye Disease Investigator Group.³² For positive and negative fusional vergence at
148 30cm and 3m, we used a horizontal prism bar (base-out for positive fusional vergence, base-in
149 for negative fusional vergence).¹⁹ To evaluate saccades, we used the specific testing
150 procedures of our clinician. In his test of saccades, participants assumed a tandem stance and
151 attempted to move only their eyes when lights appeared and disappeared (using a gap protocol
152 in which the first light disappeared before the second appeared) on the screen. The clinician
153 evaluated their performance qualitatively along three measures: quality (bad, medium, good),
154 synchronization (bad, medium, good), and saccadic correction (many, few, none).

155 *Statistical Analysis*

156 *Reliability Estimates*

157 We evaluated test-retest reliability using statistical measures for consistency and accuracy i.e.
158 the intra class correlation coefficient (ICC)⁴² and 95% limits of agreement (LoA).⁸ We considered
159 an ICC of ≤ 0.5 as poor, 0.51 - 0.74 as moderate, 0.75 - 0.89 as good, and ≥ 0.90 as excellent
160 reliability.²⁶ When multiple participant scores were the same, we used the jitter function in the R
161 software³⁴ to slightly modify the scores when plotting the results so they would appear distinct
162 from each other.

163 The LoA were calculated as recommended in the units of the scale measured.⁸ To compare LoA
164 across tests, we also standardized the scores and reported them as percentage difference $[(T1 -$
165 $T2)/\text{mean}(T1 \& T2) * 100]$. The LoA results were summarized graphically with Bland-Altman plots.⁸
166 We used the raw scale measures commonly known to clinicians for the y-axis, and report the
167 standardized version in parentheses to provide an overview of all tests.

168 Due to the limited sample size and to avoid being overly conservative in our evaluation, we
169 followed the practical solution for addressing multiple testing proposed by Saville.³⁸ Formal

Running title: Concussion

170 multiplicity correction of confidence levels was not performed but we thoroughly report all
171 statistical assessments enabling an informal *type I error* assessment by the reader.

172 *Effect of Physical State*

173 As a sensitivity analysis, we wanted to explore if a participant's change in physical state could
174 have been associated with their vision test scores. We, therefore, compared the change in
175 BVTs and saccade scores against the change in SCAT3 symptom scores using the Pearson's
176 correlation coefficient. Although we report *p*-values for these comparisons, we caution that
177 these are minimum values as we did not correct for multiple testing.

178 All statistical analyses were performed on R software 3.4.3,³⁴ plots were created using the
179 ggplot2 package.⁵⁰

180 *A priori sample size calculation*

181 Sample size calculations were done prior to the study. We used a precision-based method for
182 sample size calculations based on the ICC. Since the expected estimate for ICCs is zero, our
183 sample size calculation relied on specifying the maximum acceptable width of the confidence
184 interval for the measure of agreement. We considered the lower bound of clinical acceptability
185 to be an ICC of 0.5,⁴⁹ and expected the true ICC for repeated assessment of the different vision
186 test scores to be at least 0.75. Therefore, we believed that our estimate imprecision (95%
187 confidence interval width / 2) should not exceed 0.25. With 20 participants, under the postulated
188 assumptions, the precision of the ICC estimate was anticipated to be ± 0.20 .⁴⁹

189 *Results*

190 Of 47 potential participants identified, 26 had scheduling conflicts, one was not able to attend
191 the second visit and was excluded from the analysis, leaving 20 participants for the final
192 analysis.

Running title: Concussion

193 Demographic data for the sample analyzed is included in Table 1. Our sample included 50%
194 females with an average age of 25.5 years (SD=4.1, range=18-35), almost all university
195 educated, with 55% wearing corrective lenses with an up-to-date prescription. None of the
196 participants had any history of vision symptoms or a history of past binocular vision therapy.
197 There were six participants who had previously sustained one or two concussions, which
198 occurred 2 to 15 years prior to our study; none of these individuals reported any residual
199 concussion symptoms at the time of our study. Additionally, the three participants who had
200 reported taking selective serotonin reuptake inhibitors or anxiolytics had not received the
201 medication for at least five years as they no longer suffered from the condition.

202

Running title: Concussion

203

Table 1: Demographic and clinical characteristics of participants	
Characteristic	No. (%)
No. of participants (1 lost to follow-up)	21
Male	10 (50)
Female	10 (50)
Highest level of education attained:	
High school	1 (5)
University	19 (95)
Current status:	
Enrolled in school	10 (50)
Working	5 (25)
School & working	5 (25)
Vision correction:	
Corrective lenses	11 (55)
No correction	9 (45)
No. concussion previously sustained:	
0	14 (70)
1	3 (15)
2	3 (15)
Past medical history:	
Received medication for depression	2 (10)
Received medication for anxiety	1 (5)
Received medication for ADHD	0 (0)

Running title: Concussion

204

205 Baseline (i.e. Visit 1) scores for all tests quantitatively scored can be found in Table 2. The
206 mean scores of these tests are within range of normative scores reported in the literature (see
207 *Appendix* for more details).

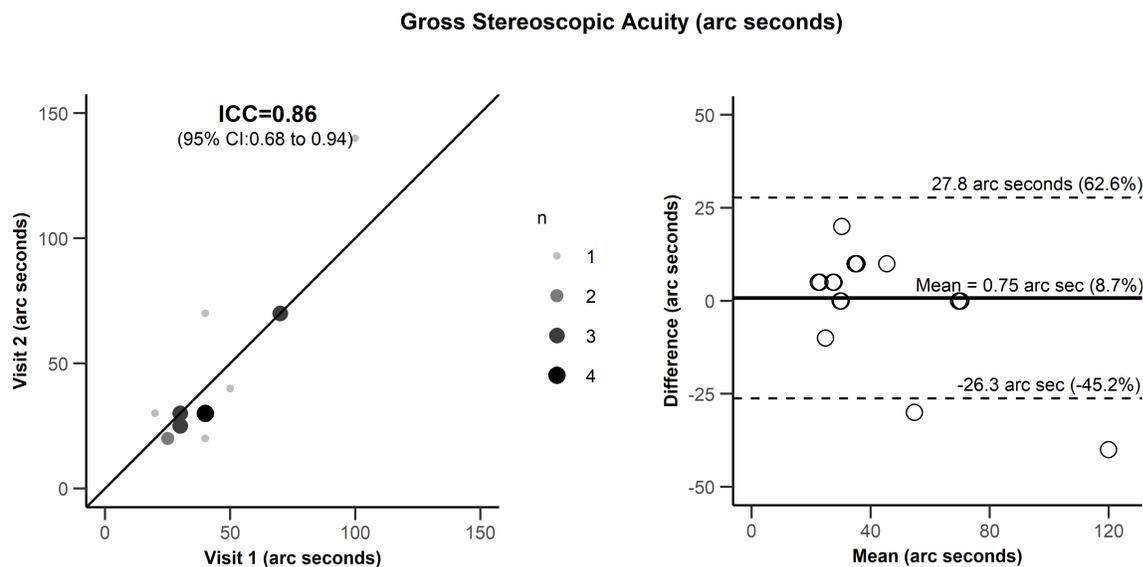
Table 2: Baseline scores for quantitatively scored vision tests	
Test (normal range)	Mean (SD)
Gross stereoscopic acuity (20-100 arc seconds)	42.5 (20.2)
Near point of convergence (3-7 cm)	4.6 (1.1)
Near point of convergence – break (1-8 cm)	5.2 (1.5)
Phoria – 3m (44-90 prism diopters)	0.3 (1.0)
Phoria – 30cm (1-45 prism diopters)	19.5 (11.2)
Positive fusional vergence – 3m (6-45 prism diopters)	16.8 (8.5)
Positive fusional vergence – 30cm (10-40 prism diopters)	25.3 (9.5)
Negative fusional vergence – 3m (2-8 prism diopters)	5.4 (1.8)
Negative fusional vergence – 30cm (4-30 prism diopters)	18.5 (6.7)

208

209 Two BVTs achieved good reliability. For phoria 3m (ICC=0.88), 18 out of 20 pairs of
210 measurements were identical at test and retest, with one participant scoring 0 and 1 prism
211 diopters, and the other scoring 0 and 2 prism diopters. Due to this highly skewed score
212 distribution, the 95% LoA are relatively uninformative (data not shown). The ICC for gross
213 stereoscopic acuity was 0.86 (Figure 1). For this test, 5 out of 20 pairs of measurements were
214 identical, with the remaining pairs differing by 5 to 40 arc seconds; the 95% LoA was ± 27.6 arc
215 seconds.

Running title: Concussion

216



217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

Figure 1: The left graph is a scatter plot for the test results at the first visit (x-axis) and retest results at the second visit (y-values) for gross stereoscopic acuity (GSA). The intra-class coefficient and its 95% confidence intervals are illustrated on the graph. The size of the gray dots (and n in the legend) represents the number of subjects with the values shown on the graph. The line of equality indicates where all points would fall if reliability was perfect. The right graph represents the Bland-Altman plot with the mean of the test-retest values on the x-axis and the difference between the test-retest results on the y-axis. The solid line represents the bias and the dotted lines represent the 95% limits of agreement (LoA). The y-axis scale represents the raw units of the test because these are the most relevant to the clinician treating the patient. Because we conducted many tests and readers may be interested in comparing the LoA across tests, we also report LoA as percent difference (T1-T2/mean of T1&T2) in parentheses.

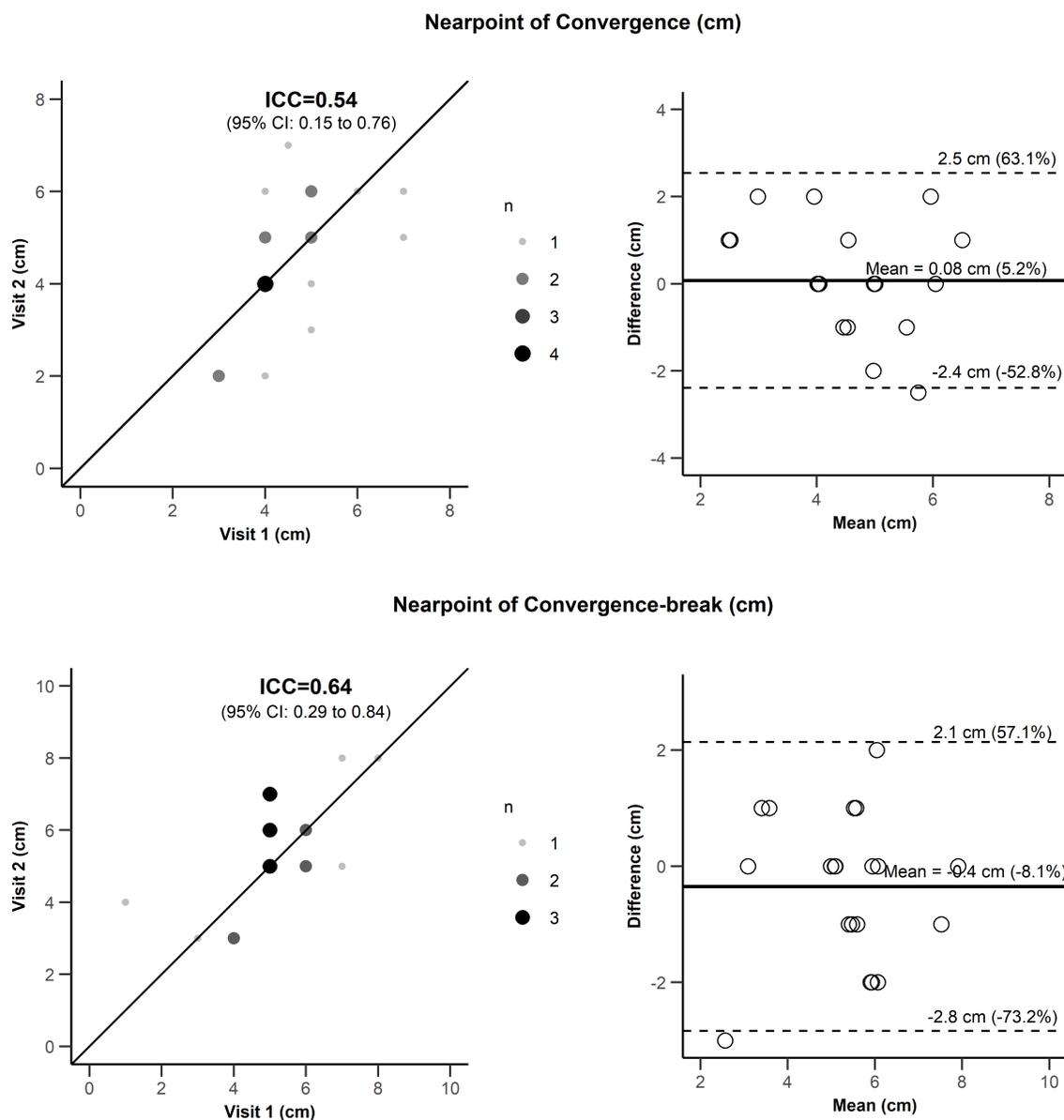
Running title: Concussion

232 We found moderate reliability for near point of convergence, near point of convergence – break,
233 phoria 30cm, and for positive and negative fusional vergence at 30cm (Figures 2 and 3), with
234 ICCs ranging between 0.54 to 0.69. For these BVTs, the 95% LoA was ± 2.5 cm for near point of
235 convergence, ± 2.5 cm for near point of convergence – break, ± 16.3 prism diopters for phoria
236 30cm, and ± 17.3 prism diopters and ± 10.4 prism diopters respectively for positive and negative
237 fusional vergence at 30cm.

238

Running title: Concussion

239



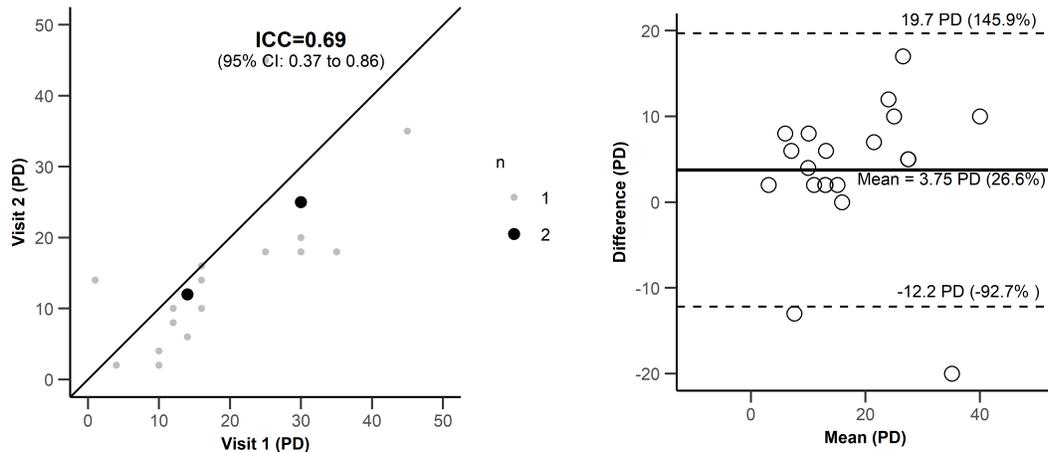
240

241 **Figure 2:** Scatter plots with intra-class coefficient results for the test-retest results
242 (left) and limits of agreement (LoA, right) for one-week test-retest reliability for 2/5
243 binocular vision tests with moderate reliability. Legends are identical to Figure 1.

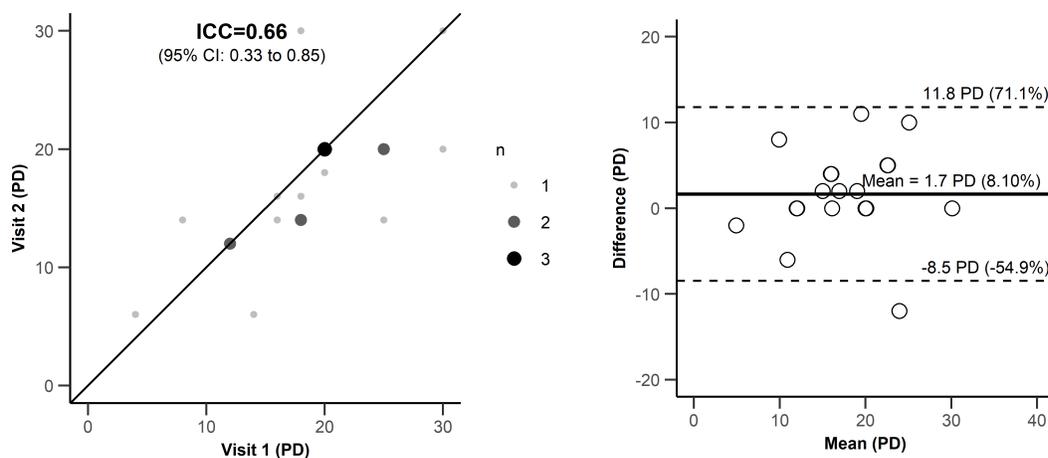
244

Running title: Concussion

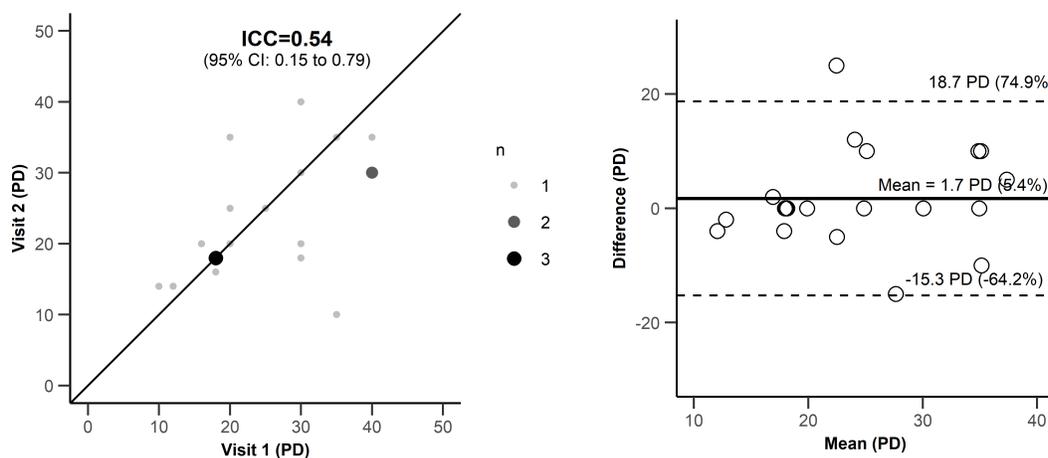
Phoria 30cm (prism diopters, PD)



Negative Fusional Vergence 30cm (prism diopters, PD)



Positive Fusional Vergence 30cm (prism diopters, PD)



Running title: Concussion

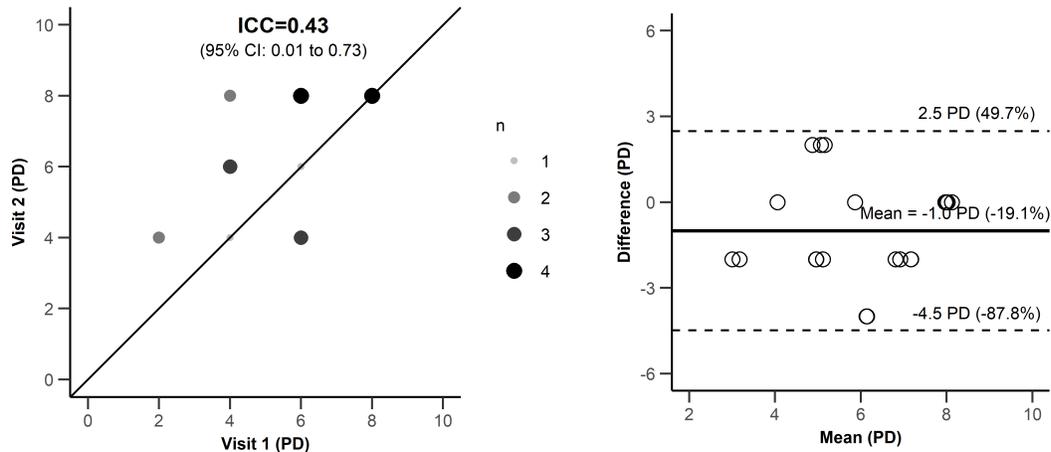
246 **Figure 3:** Scatter plots with intra-class coefficient results for the test-retest results
247 (left) and limits of agreement (LoA, right) for one-week test-retest reliability for the
248 remaining 3/5 binocular vision tests with moderate reliability. Legends are identical
249 to Figure 1.

250

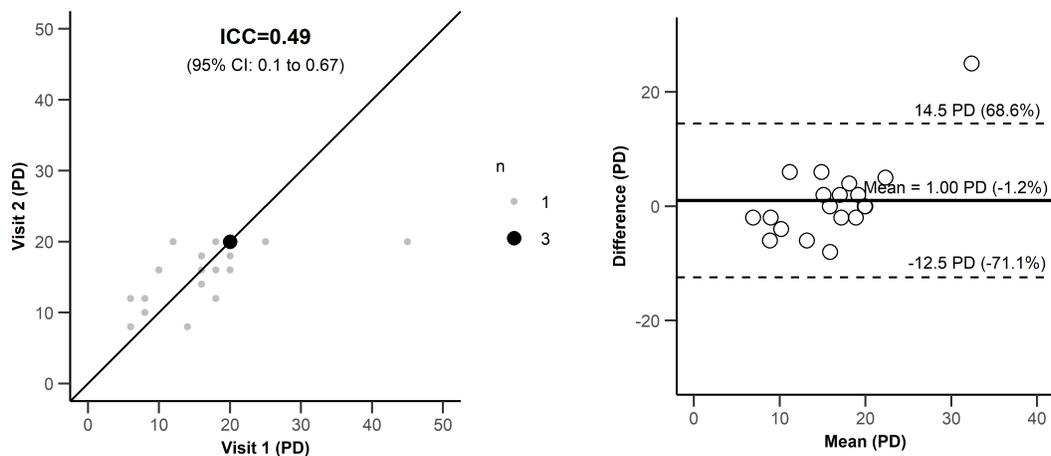
251 The three tests with poor reliability were positive and negative fusional vergence at 3m, and
252 saccades (Figure 4).

Running title: Concussion

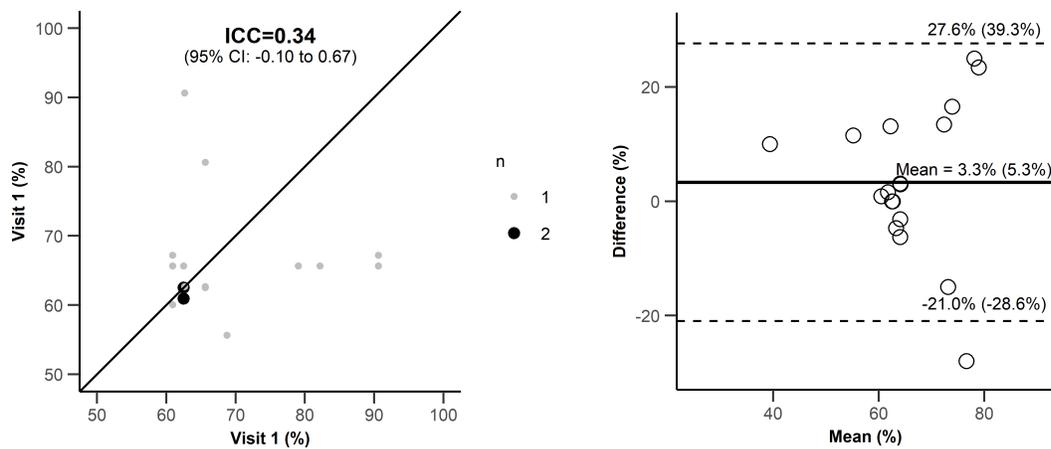
Negative Fusional Vergence 3m (prism diopters, PD)



Positive Fusional Vergence 3m (prism diopters, PD)



Saccades (%)



Running title: Concussion

254 **Figure 4:** Scatter plots with intra-class coefficient results for the test-retest results
255 (left) and limits of agreement (LoA, right) for one-week test-retest reliability for
256 saccades and the 2 binocular vision tests with poor reliability. Legends are identical
257 to Figure 1.

258

259 *Effect of Changes in Physical State*

260 In our sensitivity analysis, a participant's physical state (as measured by the SCAT3 symptom
261 score) was not relevantly associated with their BVT scores across testing sessions. Pearson's
262 correlations between changes in vision test scores and changes in symptom scores ranged
263 from -0.006 to 0.31 for all vision tests (p -values ranged from 0.19 to 0.98).

264 *Discussion*

265 Our results suggest that only 2 out of 10 vision tests demonstrated good reliability, and 5
266 additional tests had moderate reliability. There was poor reliability for saccades and both
267 positive and negative fusional vergence at 3m. The 95% LoA suggests that even with good or
268 moderate reliability, one can expect that scores for an individual with repeated measures may
269 vary by 50-70% of the mean score across all measures even if there is no change in visual
270 function. These results highlight the need for more accurate, quantifiable, and repeatable tests
271 since one might expect even more variability in a patient population compared to the healthy
272 population that we studied. Further studies are necessary to determine if changes to visual
273 function with concussion or other neurological injury (the resultant signal) are large enough to
274 be noticed given the amount of inherent noise in the tests.

275 We used the Randot Stereotest to assess gross stereoscopic acuity. Our ICC of 0.86 and 95%
276 LoA of ± 27.6 arc seconds support previous findings of good reliability in adults in the one-week

Running title: Concussion

277 time frame. The 95% LoA were previously reported as ± 0.57 *log arc seconds* (our results are
278 ± 0.58 *log arc seconds* using the same method of calculation) based on 36 patients between 7 to
279 76 years of age; time between testing intervals ranged from 10 to 364 days using a different
280 examiner at each time point.² Another study reported that 82.0% of their participants had
281 identical results at test and retest taken same day in 111 adult and children with normal vision,
282 but normal psychometric properties such as ICC or LoA were not reported.⁴⁸ A study examining
283 the one-week test-retest reliability of gross stereoscopic acuity using the related Titmus fly test
284 in 90 children reported perfect reliability (ICC=1.0)²⁸. The Random dot “E” stereotest reported
285 only interrater agreement ($K_w=0.33-0.44$) in 1257 children, but not test-retest reliability.⁴⁶

286 In the literature, there is no clear distinction between measures of near point of convergence
287 and near point of convergence – break. Further, near point of convergence is sometimes
288 referred to as a measurement of gross convergence with fusional convergence²² and other
289 times as gross convergence with proximal convergence.^{21,22,47} The inconsistent use of these
290 terms complicates comparisons across studies when the measurement procedure is not
291 reported. The Convergence Insufficiency Treatment Trial Study¹³ considered near point of
292 convergence – break as the point “[w]hen diplopia was reported,” which is consistent with our
293 definition.^{7,23} We defined near point of convergence as the closest point at which one eye
294 diverges outwards.^{9,23,37} Across the literature, near point of convergence and near point of
295 convergence – break are sometimes used interchangeably. For instance, near point of
296 convergence has been defined as the point “when the target blurs, jumps or becomes double,”³⁷
297 “when [the participant] saw 2 distinct images,”³¹ and “when the patient reported diplopia”.⁴

298 We found moderate reliability for both near point of convergence and near point of convergence
299 – break, whereas others have reported good to excellent reliability.^{31,36} In our measurement
300 procedure, participants fixated on a target that the clinician moved towards their eyes in free
301 space as used by some clinicians and researchers in the concussion field.²⁹ Others used an

Running title: Concussion

302 accommodative target, such as the Royal Air Force (RAF) rule,^{7,37} or Astron International
303 (ACR/21) Accommodative Rule.^{13,22,36,40} The RAF rule had good (ICC=0.84) test-retest reliability
304 for near point of convergence in 3 subjects with idiopathic neck pain and 7 healthy subjects, but
305 the test interval was only specified as less than one-week.¹⁸ The Astron International
306 Accommodative Rule had excellent one-week test-retest reliability (ICC=0.94-0.98) in 20
307 healthy children for near point of convergence – break.³⁶ Although tests using these
308 accommodative targets may have increased reliability compared to the methods used in this
309 study, one would generally like to minimize the accommodative load in patients with concussion
310 because it may increase symptoms. Some of the variability in our results is likely explained by
311 accommodation variability in our participants. We found no studies directly comparing the
312 reliability of the different procedures. Sheiman et al³⁹ and Rouse et al³⁶ provide a more
313 complete discussion of the advantages and disadvantages of the different methods used to
314 assess near point of convergence.

315 Measurements of phoria using the prism alternate cover test had good reliability for distance
316 (ICC=0.88) and moderate reliability for near (ICC=0.69). These results are consistent with other
317 studies which measured adult and child participants with strabismus or esotropia,^{14,32} even
318 though none of our participants had these conditions. Despite the similarity in findings, our
319 analysis methods differed slightly. For instance, because different prism increments are used to
320 measure smaller (2-20 prism diopters) or larger (>20 prism diopters) angles, other authors
321 analyzed and reported these strata separately.^{14,32} Unlike other authors, we evaluated all angles
322 of deviation together.

323 For both positive and negative fusional vergence, we reported moderate reliability (ICC=0.54,
324 0.66 respectively) for near fixation and poor reliability (ICC=0.49, 0.43 respectively) for distance
325 fixation. These results are contrary to studies reporting lower within-subject variability at near
326 fixation,³³ or no differences due to distance.⁵ Further, we found that negative fusional vergence

Running title: Concussion

327 had slightly higher reliability than positive fusional vergence at near, but were less reliable at
328 distance. However, standard clinical practice and evidence suggests the opposite; negative
329 fusional vergence is considered to have less reliability than positive fusional vergence.^{3,35} It is
330 possible that the order in which fusional vergences are taken may influence their scores. We
331 measured fusional vergences grouped by distance: (1) negative fusional vergence far, (2)
332 positive fusional vergence far, (3) negative fusional vergence near, and (4) positive fusional
333 vergence near. However, the reliability was poor for both tests at distance suggesting the order
334 of test administration would not explain the discrepancy between our results and the literature. It
335 remains possible that our results are different than others because of slight differences in our
336 methods that are not apparent in the description of the tests (see *Appendix* for full description of
337 our methods).

338 The test of saccades had the lowest ICC and poorest reliability of all the vision tests. We used
339 the clinical procedures our clinician uses in his daily practice with his patients. Participants
340 assumed a tandem stance and attempted to follow appearing and disappearing lights on a
341 screen under a gap paradigm with only their eyes, trying to keep their head still. The clinician
342 stood beside the screen in front of the participant to observe their eye movements. In some
343 published saccade test protocols, the participant's head is held still with a chin rest and a
344 forehead support to ensure that only the eyes are tracking the movements.^{16,30} In the NSUCO
345 oculomotor test, the head is not held still.⁴¹ However, unlike other tests of saccades, our
346 clinician had patients take a tandem stance which adds an additional vestibular challenge. The
347 added challenge may influence their performance on the task and introduce more noise. Finally,
348 the evaluation of the saccades was qualitative, dependent solely on the judgment of the clinician
349 with no objective measure.

Running title: Concussion

350 *Strengths and Limitations*

351 We selected a seven-day interval between testing times to evaluate the test-retest reliability of
352 the vision tests. This allowed for normal variation over time due to sleep, stress, and other
353 factors in order to provide an ICC that is applicable to following patients over time. In addition, it
354 avoids any lingering symptoms following a test that might lead to an underestimated ICC. The
355 seven-day interval also increased the likelihood the clinician remained blinded to the previous
356 results and facilitated participant recruitment because we could select a day and time when
357 participants were generally available. Some studies previously evaluated interrater
358 reliability.^{14,17,32,36} Although this has merit when one is interested in tests being evaluated by
359 more than one clinician as what might occur in group practice or a research study, interrater
360 reliability is less important when patients are followed by a single clinician over time. Our
361 objective was to define the expected “noise” when a single clinician follows a single patient over
362 time, as would occur in our target condition (i.e. concussions), so that clinicians can
363 appropriately interpret changes in the vision test scores. We provided results based on different
364 perspectives of reliability. The ICC is a measure of variability due to genuine differences in the
365 participant or due to measurement error. For instance, the ICC was 0.88 for phoria 3m,
366 indicating that 88% of the variability in the measurements was due to differences between
367 participants, and 12% was due to noise within the measurement of a participant. In addition, the
368 95% LoA provides the magnitude of the noise that can be expected with repeated measures.
369 Differences between tests at baseline and after diagnosis of a condition (e.g. concussion) likely
370 represent a true signal of a change in vision tests within the patient if these differences are
371 larger than the noise (i.e. LoA) found in our study.

372 Our study also had limitations with respect to participant population and testing measures. We
373 had a relatively small homogeneous sample size of 20 participants who were recruited via
374 convenience sampling in a university setting. However, our participants did include an equal

Running title: Concussion

375 number of males and females, over half wore corrective lenses, and the age range was 18 to 35
376 years. Our study population was thus relatively representative of our target population of young
377 athletes who may sustain concussions. The approaches used by our clinician were standard to
378 his clinical practice and were used on both of the testing days for all participants. However, the
379 methods he used to assess vision function sometimes differed from testing procedures reported
380 in the literature. For instance, a testing distance of 30 cm was used instead of the standard 40
381 cm distance for near testing of positive and negative fusional vergence.⁵ The near testing
382 distance of 30 cm was also used for phoria, which is similar to the distance commonly used in
383 the literature, 1/3 m.^{14,32} However, we are not aware of any studies comparing the effect of
384 distance on reliability. Our clinician did not attempt to separate out accommodative testing from
385 convergence (i.e. near point of convergence and near point of convergence - break, also known
386 as relative convergence) although this may be possible.¹⁵ Therefore, our measure of
387 convergence could have been affected by accommodative issues. The saccadic eye movement
388 test of our clinician also differs from commonly used tests in clinical practice and the scoring of
389 this test was qualitative and subjective, which could lead to increased variability and
390 inconsistency in scoring. Developing more quantifiable and reliable testing methods is
391 particularly important for conditions such as concussions, as they are characterized by many
392 symptoms which may only lead to subtle changes that are not detectable with imprecise tests.

393 *Conclusion*

394 We found that only 2 of 9 BVTs had good one-week test-retest reliability that could detect small
395 to moderate changes in visual function, and an additional 5 BVTs that might be able to detect
396 moderate change in visual function. The remaining two BVTs and saccades may still be useful if
397 changes in visual function are expected to be larger than the noise of the measure.

398

Running title: Concussion

399 *Acknowledgements*

400 We would like to thank Isabel Pereira for her help throughout the course of this work. We would
401 also like to thank David Tinjust, the clinician who conducted these tests in our participant
402 population and who provided partial funding for this study. This study was funded through
403 programs designed to foster collaboration between academics and industry. The government
404 sources are the MITACS program, and the MEDTEQ program. The industry partners were
405 Apexk Inc, Varitron Technologies Inc and l'Institut National du Sport du Quebec.

406 *Appendix*

407 The Appendix expands on the concise description of the nine binocular vision tests and test of
408 saccades provided in the manuscript. This is especially important when comparing results
409 across different studies in evidence synthesis.

410

Running title: Concussion

411 *References*

- 412 1. SCAT3. *Br J Sports Med*. 2013;47(5):259.
- 413 2. Adams WE, Leske DA, Hatt SR, Holmes JM. Defining real change in measures of stereoacuity.
414 *Ophthalmology*. 2009;116(2):281-285.
- 415 3. Alpern M. The after effect of lateral duction testing on subsequent phoria measurements. *Am J*
416 *Optom Arch Am Acad Optom*. 1946;23(10):442-447.
- 417 4. Alvarez TL, Kim EH, Vicci VR, Dhar SK, Biswal BB, Barrett AM. Concurrent vision dysfunctions in
418 convergence insufficiency with traumatic brain injury. *Optometry and vision science : official*
419 *publication of the American Academy of Optometry*. 2012;89(12):1740-1751.
- 420 5. Antona B, Barrio A, Barra F, Gonzalez E, Sanchez I. Repeatability and agreement in the
421 measurement of horizontal fusional vergences. *Ophthalmic & physiological optics : the journal of*
422 *the British College of Ophthalmic Opticians (Optometrists)*. 2008;28(5):475-491.
- 423 6. Biousse V, Skibell B, Watts R, Loupe D, Drews-Botsch C, Newman N. Ophthalmologic features of
424 Parkinson's disease. *Neurology*. 2004;62:177-180.
- 425 7. Bishop A. Convergence and convergent fusional reserves - investigation and treatment. In: Doshi
426 S, Evans BJW, eds. *Binocular Vision and Orthoptics: Investigation and Management*. Oxford:
427 Butterworth-Heinemann; 2001:30-62.
- 428 8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of
429 clinical measurement. *Lancet*. 1986;1(8476):307-310.
- 430 9. Bredemeyer H, Bullock K. Methods of examination. *Orthoptics: Theory and Practice*. St. Louis:
431 The C.V. Mosby Company; 1968:130-152.
- 432 10. Brozek J, Simonson E, Bushard W, Peterson J. Effects of practice and the consistency of repeated
433 measurements of accommodation and vergence. *Am J Ophthalmol*. 1948;31(2):191-198.

Running title: Concussion

- 434 **11.** Capo-Aponte JE, Urosevich T, Temme L, Tarbett A, Sanghera N. Visual dysfunctions and
435 symptoms during the subacute stage of blast-induced mild traumatic brain injury. *Mil Med.*
436 2012;177:804-815.
- 437 **12.** Ciuffreda KJ, Kapoor N, Rutner D, Suchoff IB, Han ME, Craig S. Occurrence of oculomotor
438 dysfunctions in acquired brain injury: a retrospective analysis. *Optometry (St Louis, Mo).*
439 2007;78(4):155-161.
- 440 **13.** Convergence Insufficiency Treatment Trial Study G. The convergence insufficiency treatment
441 trial: design, methods, and baseline data. *Ophthalmic Epidemiol.* 2008;15(1):24-36.
- 442 **14.** de Jongh E, Leach C, Tjon-Fo-Sang MJ, Bjerre A. Inter-examiner variability and agreement of the
443 alternate prism cover test (APCT) measurements of strabismus performed by 4 examiners.
444 *Strabismus.* 2014;22(4):158-166.
- 445 **15.** Digre KB. Principles and techniques of examination of the pupils, accommodation, and
446 lacrimation. In: Miller NR, Newman NJ, Biousse V, Kerrison JB, eds. *Walsh and Hoyt's Clinical*
447 *Neuro-Ophthalmology.* 6th ed. Philadelphia: Lippencott Williams & Wilkins; 2005:715-738.
- 448 **16.** Ettinger U, Kumari V, Crawford T, Davis R, Sharma T, Corr P. Reliability of smooth pursuit,
449 fixation, and saccadic eye movements. *Psychophysiology.* 2003;40:60-628.
- 450 **17.** Fawcett SL, Birch EE. Interobserver test-retest reliability of the Randot preschool stereoacuity
451 test. *Journal of AAPOS : the official publication of the American Association for Pediatric*
452 *Ophthalmology and Strabismus.* 2000;4(6):354-358.
- 453 **18.** Giffard P, Daly L, Treleaven J. Influence of neck torsion on near point convergence in subjects
454 with idiopathic neck pain. *Musculoskelet Sci Pract.* 2017;32:51-56.
- 455 **19.** Goss DA, Becker E. Comparison of near fusional vergence ranges with rotary prisms and with
456 prism bars. *Optometry (St Louis, Mo).* 2011;82(2):104-107.

Running title: Concussion

- 457 **20.** Granet D, Gomi C, Miller-Scholte A. The relationship between convergence insufficiency and
458 ADHD. *Strabismus*. 2005;13:163-168.
- 459 **21.** Grosvenor T. *Primary Care Optometry: Anomalies of Refraction and Binocular Vision*. Boston:
460 Butterworth-Heinemann; 1996.
- 461 **22.** Hayes GJ, Cohen BE, Rouse MW, De Land PN. Normative values for the nearpoint of
462 convergence of elementary schoolchildren. *Optometry and vision science : official publication of*
463 *the American Academy of Optometry*. 1998;75(7):506-512.
- 464 **23.** Hurtt J, Rasicovici A, Windsor C. Diagnosis and diagnostic tests. *Comprehensive Review of*
465 *Orthoptics and Ocular Motility: Theory, Therapy, and Surgery/*. 2nd ed. Saint Louis: The C.V.
466 Mosby Company; 1977:123-152.
- 467 **24.** Kaas JH. The evolution of the complex sensory and motor systems of the human brain. *Brain Res*
468 *Bull*. 2008;75(2-4):384-390.
- 469 **25.** Kontos AP, Elbin RJ, Schatz P, et al. A revised factor structure for the post-concussion symptom
470 scale: baseline and postconcussion factors. *Am J Sports Med*. 2012;40(10):2375-2384.
- 471 **26.** Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for
472 reliability research. *J Chiropr Med*. 2016;15(2):155-163.
- 473 **27.** Maples WC, Hoenes R. Near point of convergence norms measured in elementary school
474 children. *Optom Vis Sci*. 2007;84(3):224-228.
- 475 **28.** Moganeswari D, Thomas J, Srinivasan K, Jacob GP. Test re-test reliability and validity of different
476 visual acuity and stereoacuity charts used in preschool children. *J Clin Diagn Res*.
477 2015;9(11):NC01-05.
- 478 **29.** Mucha A, Collins MW, Elbin RJ, et al. A brief vestibular/ocular motor screening (VOMS)
479 assessment to evaluate concussions: preliminary findings. *Am J Sports Med*. 2014;42(10):2479-
480 2486.

Running title: Concussion

- 481 **30.** Paut O, Vercher J-L, Blin O, et al. Evaluation of saccadic eye movements as an objective test of
482 recovery from anaesthesia. *Acta Anaesthesiol Scand*. 1995;39:1117-1124.
- 483 **31.** Pearce KL, Sufrinko A, Lau BC, Henry L, Collins MW, Kontos AP. Near point of convergence after
484 a sport-related concussion: Measurement reliability and relationship to neurocognitive
485 impairment and symptoms. *Am J Sports Med*. 2015;43(12):3055-3061.
- 486 **32.** Pediatric Eye Disease Investigator Group. Interobserver reliability of the prism and alternate
487 cover test in children with esotropia. *Arch Ophthalmol*. 2009;127(1):59-65.
- 488 **33.** Penisten D, Hofstetter H. Reliability of rotary prism fusional vergences ranges. *Optometry (St*
489 *Louis, Mo)*. 2001;72:117-122.
- 490 **34.** R Core Team. R: a language and environment for statistical computing. . Vienna, Austria: R
491 Foundation for Statistical Computing; 2015.
- 492 **35.** Rosenfield M, Ciuffreda KJ, Ong E, Super S. Vergence adaptation and the order of clinical
493 vergence range testing. *Optometry and vision science : official publication of the American*
494 *Academy of Optometry*. 1995;72(4):219-223.
- 495 **36.** Rouse M, Borsting E, Deland P, The Convergence Insufficiency and Reading Study (CIRS) Group.
496 Reliability of binocular vision measurements used in the classification of convergence
497 insufficiency. *Optometry and vision science : official publication of the American Academy of*
498 *Optometry*. 2002;79(4):254-264.
- 499 **37.** Rowe F. Investigative procedures. *Clinical Orthoptics*. 3rd ed. Chichester, West Sussex: Wiley-
500 Blackwell; 2012:62-78.
- 501 **38.** Saville DJ. Multiple comparison procedures: The practical solution. *Am Stat*. 1990;44(2):174-180.
- 502 **39.** Scheiman M, Gallaway M, Frantz KA, et al. Nearpoint of convergence: test procedure, target
503 selection, and normative data. *Optometry and vision science : official publication of the*
504 *American Academy of Optometry*. 2003;80(3):214-225.

Running title: Concussion

- 505 **40.** Scheiman M, Mitchell GL, Cotter S, et al. A randomized clinical trial of treatments for
506 convergence insufficiency in children. *Arch Ophthalmol.* 2005;123:14-24.
- 507 **41.** Scheiman M, Wick B. Diagnostic testing. *Clinical management of binocular vision: Heterophoric*
508 *accommodative, and eye movement disorders.* Maryland, USA: Wolters Kluwer/Lippincott
509 Williams & Wilkins; 2015:3-49.
- 510 **42.** Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological*
511 *bulletin.* 1979;86(2):420-428.
- 512 **43.** Stereo Optical Co. Randot Stereotests. In: Stereo Optical Co., ed; 1995.
- 513 **44.** Suchoff IB, Kapoor N, Ciuffreda KJ. An overview of acquired brain injury and optometric
514 implications. In: Foundation OEP, ed. *Visual and vestibular consequences of acquired brain*
515 *injury.* Santa Ana, CA: Optometric Extension Program Foundation; 2001.
- 516 **45.** Ventura RE, Balcer LJ, Galetta SL. The Concussion Toolbox: The Role of Vision in the Assessment
517 of Concussion. *Semin Neurol.* 2015;35(5):599-606.
- 518 **46.** Vision in Preschoolers (VIP) Study Group. Random Dot E Stereotest: Testability and reliability in
519 3- to 5-year old children. *Journal of AAPOS : the official publication of the American Association*
520 *for Pediatric Ophthalmology and Strabismus.* 2006;10(6):507-514.
- 521 **47.** von Noorden G. Examination of the patient-II: motor signs in heterophoria and heterotropia.
522 *Binocular Vision Motility and Ocular Motility.* St. Louis: Mosby; 2002:206-207.
- 523 **48.** Wang J, Hatt SR, O'Connor AR, et al. Final version of the distance Randot Stereotest: normative
524 data, reliability, and validity. *Journal of AAPOS : the official publication of the American*
525 *Association for Pediatric Ophthalmology and Strabismus.* 2010;14(2):142-146.
- 526 **49.** Watson PF, Petrie A. Method agreement analysis: a review of correct methodology.
527 *Theriogenology.* 2010;73(9):1167-1179.
- 528 **50.** Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2009.

Running title: Concussion

- 529 **51.** Yengo-Kahn A, Hale A, Zalneraitis B, Zuckerman SL, Sills A, Soloman G. The sports concussion
530 assessment tool: a systematic review. *Neurosurg Focus*. 2016;40(4):E6.
- 531