







Original Research

Feasibility and Reliability of a Novel Game-Based Test of Neurological Function in Youth: The Equilibrium Test Battery

Heather A Shepherd¹ ^a, Cody R van Rassel² , Amanda M Black¹ , Robert F Graham³, Keith Owen Yeates⁴ ,
Carolyn A Emery⁵ , Kathryn J Schneider⁶ 

¹ Sport Injury Prevention Research Centre, Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada; Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada; Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada; O'Brien Institute for Public Health, University of Calgary, Calgary, AB, Canada, ² Sport Injury Prevention Research Centre, Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada; Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada; Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada; O'Brien Institute for Public Health, University of Calgary, Calgary, AB, Canada; Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada, ³ Sport Injury Prevention Research Centre, Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada, ⁴ Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada; Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada; Department of Psychology, University of Calgary, Calgary, AB, Canada, ⁵ Sport Injury Prevention Research Centre, Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada; Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada; Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada; O'Brien Institute for Public Health, University of Calgary, Calgary, AB, Canada; Pediatrics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; Community Health Sciences, Department of Pediatrics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, ⁶ Sport Injury Prevention Research Centre, Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada; Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada; Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada; Sport Medicine Centre, Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada; Evidence Sport and Spine, Calgary, AB, Canada

Keywords: assessment, concussion, digital technology, feasibility, adolescents

<https://doi.org/10.26603/001c.32527>

International Journal of Sports Physical Therapy

Vol. 17, Issue 3, 2022

Background

An estimated 11% of Canadian adolescents will sustain a sport-related concussion each year. However, diagnostic tools to detect and monitor concussive outcomes are limited.

Purpose

To evaluate the feasibility and test-retest reliability of the Highmark Interactive Equilibrium (HIEQ) test battery in uninjured adolescents.

Study Design

Observational study with repeated measurements.

Methods

Participants completed the HIEQ test battery, a game-based platform on an iPad application, that assesses balance, cognitive function, and visual function, for up to 15 consecutive school days in a group classroom setting. Feasibility for use of the HIEQ was evaluated by (1) recruitment rates; (2) retention rates; (3) test completion without assistance; and (4) adverse events. Test-retest reliability was examined using Bland Altman 95% limits of agreement and intraclass correlation coefficients comparing the first and second and second and third obtained scores. Reliability across multiple baseline assessments was also analyzed using intraclass correlations for the second to sixth and seventh to eleventh obtained scores.

a Corresponding author:

Heather Shepherd
Heather.shepherd@ucalgary.ca
403.220.5332
Sport Injury Prevention Research Centre,
Faculty of Kinesiology, University of Calgary
Calgary, AB
T2N 1N4
Canada

Results

Fifty-five uninjured high school students (31 females and 24 males, mean age = 16.24 [1.09]) from three high schools participated in the study. Three participants (5%) completed all 15 days of testing, and 73% completed at least 10 of 15 test days. No adverse events were reported. Although the test was feasible, all subtests showed wide limits of agreement from first to second and from second to third testing occasions. Results indicate poor-to-moderate reliability (<0.50 to 0.75) across those intervals, as well as across the second to sixth and seventh to eleventh testing occasions.

Conclusion

The HIEQ is feasible in high school students; however, performance was characterized by wide limits of agreement and poor-to-moderate reliability across test occasions. Future evaluation of the HIEQ in visual and auditory distraction free individual testing settings is warranted.

Level of Evidence

Level 3.

INTRODUCTION

In Canada, approximately one in nine adolescents will sustain a sport-related concussion each year.¹ Concussion signs and symptoms are heterogenous and may involve a variety of domains, including physical (e.g. headache, dizziness), emotional (e.g. irritability, anxiety), sleep, and cognitive (e.g. difficulty remembering, difficulty concentrating).² Concussion assessments should be multifaceted, incorporating multiple domains of functioning.³

Once a concussion has been diagnosed, decisions regarding recovery and clearance to return to school or to sport are made on an individual, case-by-case basis. An objective tool that assesses multiple domains at baseline and throughout the recovery process may help make informed diagnostic, management, and recovery decisions for individuals who sustain a concussion.

Current computerized neuropsychological assessment batteries used for concussion assessment (e.g., ImpACT, Axon Sports CogState Test) demonstrate variable levels of reliability,^{4,5} which may be attributed, in part, to natural variation in an individual's performance on a given performance task. However, variable performance may reduce reliability and limit a test battery's clinical utility.⁶ This has resulted in a shift towards conducting multiple baselines, in order to better understand normative variation in performance across multiple domains of function.⁶ Multiple baselines may help to yield a more reliable estimate of baseline performance, and may support clinicians in detecting change in an athlete's functioning following a concussion.⁶

The Highmark Interactive Equilibrium (HIEQ) test battery is a game-based platform that draws on existing clinical tools, including the Sport Concussion Assessment Tool (SCAT), King-Devick Test, and Trail Making Test Part B.⁷⁻¹¹ The HIEQ assesses multiple domains of functioning including balance, vision, and cognition (i.e., immediate verbal recognition memory, delayed verbal recognition memory, cognitive flexibility and inhibitory control, and attention and working memory).¹² The HIEQ is unique in that it is self-administered by the participant, as opposed to requiring the expertise of a clinician to administer, and hence needs minimal supervision.¹³

The feasibility of the HIEQ, as well as its test-retest reliability, has not been studied extensively in healthy adolescents. Therefore, the purpose of this study was to evaluate the feasibility and test-retest reliability of the HIEQ test battery in uninjured adolescents.

MATERIALS AND METHODS

STUDY DESIGN AND PARTICIPANTS

This was an observational study that involved repeated measurements in a sample of high school students. Students enrolled in three high schools in Calgary, Alberta, Canada were invited to participate in the study. Recruitment initially occurred at the school level with school administrators. Once administrators agreed to allow the research team to approach teachers within a school, students enrolled in sport medicine, sports performance, and biology classes were invited to participate in the study. Two of the schools enrolled students who were focused on sport, including athletes participating at highly competitive levels. Some of the students were members of National teams or other high-performance levels of sport. The third school was a regular public high school.

Participants were eligible for inclusion if they met the following criteria: 1) age 14-19 years; 2) high school student in one of the participating classes; 3) uninjured at the start of the study; and 4) provided written participant and/or guardian informed consent and participant assent. Participants were excluded if they had any musculoskeletal injuries or other medical conditions that would hinder completion of the test battery, or if they had any planned absences from school during the study period that would preclude completion of the test battery daily for three consecutive weeks.

PROCEDURES

This study was approved by University of Calgary Conjoint Health Research Ethics Board (REB18-1482) and participating schools. Data were collected over a three-week period in each school (April-June 2019). Participants were provided detailed instructions and a demonstration of the HIEQ ap-

plication test battery (EQ Active, v. 1.1.3, Highmark Interactive, Toronto, ON, Canada), as well as instruction on how to complete the baseline questionnaire.

Participants were asked to complete the HIEQ once per school day for 15 school days. They completed the testing in their classroom under the supervision of the teacher during class time. Schools were provided with a set of iPads, managed by the teachers. Each participant was provided a unique study e-mail address to login to the application and wired headphones to reduce distractions and to minimize delays between sounds from the application and participant response times (i.e. to reduce the delay in auditory stimulus that may occur with wireless/Bluetooth headphones). Members of the research team followed up with the sites once weekly to ensure that the test was being completed and to answer any questions from students or teachers at the schools. The research team also followed up with the teachers between visits via e-mail, as needed.

The HIEQ test battery required an internet connection to complete and send the data to a secure third-party online server (Cloud66, San Francisco, CA, US) for data storage. MongoDB Compass (Mongo DB, Inc., New York City, NY, USA) and Sequel Pro (v1.1.2, Sequelpro.com) were used to access data stored on the Cloud66 sever. Access to the HIEQ test battery data was limited to the research team. Highmark Interactive was blinded to the test data from the HIEQ application throughout the study.

MATERIALS

BASELINE QUESTIONNAIRE

The baseline questionnaire included questions pertaining to demographics, history of concussion, history of other injuries, and history of learning disabilities or participant-identified learning concerns. Participants completed the baseline questionnaire at the onset of study participation via a survey link through Research Electronic Data Capture (REDCap).

HIGHMARK INTERACTIVE EQUILIBRIUM APPLICATION

The HIEQ application contained seven subtests. Details on each subtest and scoring is described below. Each participant self-administered the HIEQ application and was instructed to complete the “Full Check-In”, or all seven subtests, on each testing day. The test battery took approximately 20 minutes to complete initially, but only 10-15 minutes once participants became familiar with the test.

HIGHMARK EQ APPLICATION – SUBTESTS

VISUAL FUNCTION

The visual function subtest, “Dance Off,” was designed to assess binocular vision and visual processing using a rapid, timed task. It was developed to be analogous to the King-Devick Test.^{10,11} The subtest consisted of a randomized presentation of arrows on the screen of the device that were to be scanned visually from left to right. The participant indicated the direction of each arrow by swiping in the di-

rection of the arrow on the device. Participants were asked to complete each card (trial) as quickly as possible, without incurring any errors. If an error occurred, the participant started the card again. The task involves three cards increasing in difficulty (i.e., changes in spacing between arrows). Time to complete each card was recorded in seconds. Completion times were averaged to yield a score in seconds. Higher scores indicated worse performance.

BALANCE

The balance subtest, “Tire Toss,” was designed to capture balance in five different positions: 1) Romberg (feet together); 2) Tandem stance /sharpened Romberg (right leg forward); 3) Tandem stance/sharpened Romberg (left leg forward); 4) single leg stance (right leg); and 5) single leg stance (left leg).^{14,15} The positions are the same as those used in the Balance Error Scoring System (BESS) that is part of the SCAT.^{7,16} Participants were instructed to keep their eyes closed and to keep the device (iPad) level and flat against their chest. Participants held each position for 10 seconds. The balance subtest used the internal accelerometer, which measures the orientation of the iPad with respect to the participant’s initial position, to determine a participant’s tilt, or deviation from the centre of mass, to a maximum deviation of 15° from centre. A balance score was calculated using the cumulative tilt of the iPad as an approximation of the movement of the participant’s centre of mass during the test, averaged across the five balance positions. Scores range from 0 to 100, with 100 indicating better balance, or a lower deviation from the centre of mass, and 0 a higher deviation from centre of mass. If a participant deviated 15° or more from centre, then they were assigned a score of 0 for that position. If a participant lost their balance, they were instructed to resume the initial position as soon as safely able.

COGNITIVE FUNCTION

Cognitive function was measured using five tests, four of which are similar to components of the Standardized Assessment of Concussion, which is a widely accepted screening test of concussion that is part of the SCAT.^{7,8} The five tests are named Recipe Recall, Recipe Recall with Delay, Jersey Reversey, Fast Ball, and Pylon Pivot.

“Recipe Recall” evaluated immediate verbal recognition memory, similar to immediate memory on the SCAT5.⁷ Participants were presented with an auditory list of 20 words (foods) as part of a grocery list and were instructed to remember the grocery items. They were presented with this list twice. Then, they were presented with 40 words, 20 of which were previously heard and 20 novel words, in random order. Participants selected “yes” or “no” based on whether the presented word was on the grocery list. Scores were calculated as correct out of a total of 40. Higher scores indicated better performance.

“Recipe Recall with Delay” evaluated delayed verbal recognition memory, similar to delayed recall on the SCAT5.⁷ This subtest included a similar presentation to Recipe Recall. Participants selected “yes” or “no” based on whether the presented word was on the initial grocery list

(Recipe Recall). Scores were calculated as correct out of a total of 20. Higher scores indicated better performance.

“Jersey Reversey” evaluated working memory and attention, similar to Digits Backwards on the SCAT5.⁷ Participants were presented with an auditory list of single-digit (1-9) numbers and were instructed to reverse the presented numbers. They selected the intended numbers presented on the screen. The subtest started with three numbers and progressed to a maximum of nine numbers. If participants were successful on a given level, then they progressed to the next level, increasing the number of digits by one. Participants had two trials to complete each level. If they were incorrect on both trials at a given level, then the subtest ended. Presentation of numbers was randomized. Scores were calculated as successful level achieved, out of a maximum score of seven (i.e., highest level completed). Higher scores on this subtest indicated better performance.

“Fastball” evaluated reaction time. It included five trials where participants tapped the screen as soon as a stimulus (a baseball) appeared. The timing between stimuli presentations was randomized, and stimuli could occur from one-to-three seconds apart. The subtest was scored by averaging the reaction time of the five trials, measured in milliseconds. Higher reaction time scores were indicative of worse performance. Reaction times below 100 milliseconds were considered invalid and excluded from the average. If a participant missed a trial, they were assigned a score of 1500 milliseconds, which was included in the average of the five trials.

The final subtest, “Pylon Pivot,” is analogous to the Trail Making Test Part B and evaluates inhibitory control and flexibility.⁹ Participants completed the subtest by selecting a series of numbers and letters in ascending order (e.g. 1-A, 2-B, 3-C). The locations of letters and numbers were randomized. Scoring was calculated as time to completion in seconds. If a participant made an error, they received haptic feedback and then selected the correct next number or letter in the series before progressing. Higher time scores indicated worse performance.

STATISTICAL ANALYSIS

The feasibility of the test battery was evaluated using the following metrics: 1) recruitment rates; 2) retention rates; 3) completion of test without assistance; and 4) adverse events reported. Boxplots were completed for each subtest to demonstrate medians and ranges of scores across testing days.

TEST-RETEST RELIABILITY

Test-retest reliability was assessed using Bland Altman 95% limits of agreement. Comparisons were conducted for the first versus second score and the second versus third score. Scores were adjusted for start time (i.e. if a participant’s first day of test administration was testing day 2, then that score [and all subsequent scores] would be adjusted – day 2 would equal score 1). This procedure was used to calculate missing data points for variables on Days 1, 2, and 3, with adjusted days varying up to five days (i.e., Score 1 could be from Day 1, 2, 3, 4, or 5). The Shapiro-Wilk test of normality

was used to check for the assumption of normality of difference scores for each subtest comparison. Where assumptions of normality were not met, data were logarithmically transformed. Ratios, geometric means, and 95% confidence intervals were calculated for back-transformed logarithmic data. Intraclass correlations (ICCs) and 95% confidence intervals also were used to calculate test-retest reliability for scores 1-2 and scores 2-3 based on a mean rating ($k=2$), two-way mixed-effects model with absolute agreement. Scores 1-2 and scores 2-3 were selected to consider initial test familiarization (scores 1-2) and to determine if test-retest reliability differs after the first two test administrations (scores 2-3). ICC values of less than 0.50 were categorized as poor, 0.50-0.75 as moderate, 0.76-0.90 as good, and above 0.90 as excellent.¹⁷

To evaluate test-retest reliability across multiple baseline periods, ICCs and 95% confidence intervals were calculated for intra-individual scores 2 to 6 and scores 7 to 11 based on a mean rating ($k=5$), two-way mixed-effects model with absolute agreement. Score bands 2 to 6 and 7 to 11 were used as five-day multiple baselines as recommended clinically. Score 1 was not included to minimize any potential learning effects from the first-to-second test administration. The Shapiro-Wilk test of normality was used to check for the assumption of normality of difference scores for each subtest comparison. All analyses were performed using Stata v. 15.0 and Microsoft Excel.

RESULTS

Participants included 55 14–19-year-old high school students (mean = 16.24 [1.09]; 31 females and 24 males), ranging from grades 10 through 12. Twenty-eight of the participants had a prior history of concussion, and 27 had a prior history of a non-concussive injury. Of those with prior injury history, 15 participants reported a prior history of both a concussion and a non-concussive injury. Sixteen of the students self-reported a history of learning disabilities or participant-identified learning concerns. Demographic characteristics are presented [Table 1](#).

FEASIBILITY

Three participants (5%) completed the HIEQ on all test dates. Most participants ($n = 40$; 73%) completed the battery for at least 10 of 15 testing days. The number of participants who completed the HIEQ test battery decreased as the number of testing days increased. All participants were able to complete the HIEQ test battery without assistance, and no adverse events were reported. Due to unexpected events, including school closures and loss of internet service, participation numbers were especially low on Days 8 and 9. [Figure 1](#) outlines the proportion of participants who completed a given number of testing days, out of a possible 15 days. [Figure 2](#) shows the ranges of scores for each subtest across each testing day.

TEST-RETEST RELIABILITY

Bland Altman 95% limits of agreement are presented in [Figures 3](#) and [4](#) and [Table 2](#). Plots for Recipe Recall (score 1

Table 1. Participant Demographics

	Females (n=31)	Males (n=24)
Age (years), Standard deviation	Mean = 16.13 (1.02)	Mean = 16.38 (1.17)
Previous history of concussion	Yes = 14 (45%) No = 15 (48%) Missing = 2 (7%)	Yes = 14 (58%) No = 8 (33%) Missing = 2 (9%)
Previous history of injury other than concussion	Yes = 18 (58%) No = 11 (35%) Missing = 2 (7%)	Yes = 9 (38%) No = 12 (50%) Missing = 3 (12%)

to score 2, score 2 to score 3) and Pylon Pivot (score 1 to score 2, score 2 to score 3) were logarithmically transformed to improve dispersion of scores (in which case ratios are reported). Overall, wide 95% limits of agreement and ratios were apparent across all testing points. Timed subtests (Dance Off, Pylon Pivot, Fast Ball) showed improved performance, or a decrease in mean scores, from score 1 to score 2, with two (Pylon Pivot and Fast Ball) also showing improvement from score 2 to score 3.

ICCs were calculated for test-retest reliability for score 1 to score 2 and score 2 to score 3 across subtests. Reliability ranged from poor-to-moderate, where ICCs were calculated (see Table 3).¹⁷ For score 1 to score 2, moderate test-retest reliability was found for Jersey Reversey (ICC 0.52) and Tire Toss (ICC 0.61), and poor test-retest reliability for Recipe Recall (ICC 0.18). For score 2 to score 3, moderate test-retest reliability was observed for Pylon Pivot (ICC 0.64), Jersey Reversey (ICC 0.68, Tire Toss (ICC 0.57), and Fast Ball (ICC 0.66). For score 2 to score 3, poor test-retest reliability was observed for Recipe Recall score 2 to score 3 (ICC 0.44). ICCs were not reported for score 1 to score 2 for Dance Off, Pylon Pivot, Recipe Recall with Delay, or Fast Ball, or for score 2 to score 3 for Dance Off or Recipe Recall with Delay, as normality assumptions were violated.

ICCs for scores 2 through 6 and scores 7 through 11 are presented in Table 4. ICCs ranged from 0.34 to 0.55 for score 2 to score 6 and from 0.33 to 0.53 for score 7 to score 11, or from poor-to-moderate for both intervals. ICCs were not reported where assumptions were violated.

DISCUSSION

The purpose of the current study was to evaluate the feasibility and test-retest reliability of the HIEQ application, a game-based test battery designed to assess neurological functioning, including balance, visual function, and cognition, in adolescents.

FEASIBILITY

Participants were recruited from four classes in three Calgary, Alberta, Canada high schools, with most students in the eligible classes electing to participate in the study. Because the sample included students who attended high-performance sports programs (n=29), some participants were absent from school for large portions of the data collection period. Thus, none of the subtests had 100% participation on any given testing day. This is highlighted in Figure 1, where four participants completed five or less days of test-

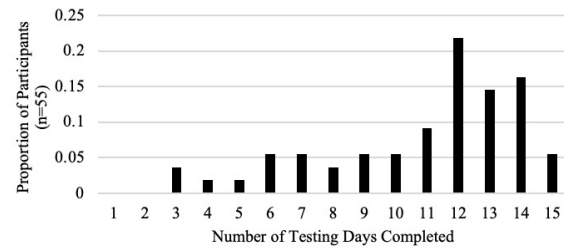


Figure 1. Number of testing days completed by participants.

ing, and only three participants completed all 15 days. Nevertheless, despite these circumstances, most participants completed the battery for at least 10 of 15 testing days. In contrast to other computer-based assessments (e.g. IMPACT) or pen-and-paper assessments (e.g. SCAT5), the HIEQ does not need a health care professional to be present during administration. Instead, adolescents can complete the assessment independently. This may increase the likelihood of repeat testing that enables multiple baseline assessments.

To the authors' knowledge, this was the first study that has evaluated multiple baselines across 15 days. Anecdotal reports from participants and teachers suggested that participants found the application engaging initially, with interest decreasing after the initial one-to-two weeks of test administration. This was also seen when examining participation rates, which were highest at the start of the study and declined after testing day seven, suggesting a possible limit to the desirable number of baseline test administrations. Participant fatigue has been identified as a factor which may affect reliability in studies using computerized neurocognitive tests, similar to HIEQ.⁴ However, this was noted to be more of a concern when multiple computerized neurocognitive tests were conducted concurrently, not specific to one test conducted across multiple days.

RELIABILITY

It was anticipated that the most improvement in mean test performance would occur from score 1 to score 2, with less improvement from score 2 to score 3. Descriptively, improvement in test performance was observed from score 1 to score 2 for timed subtests (Dance Off, Pylon Pivot, Fast Ball) and Jersey Reversey and Tire Toss. Box plots (Figure 2) were used to describe test performance over days, and sug-

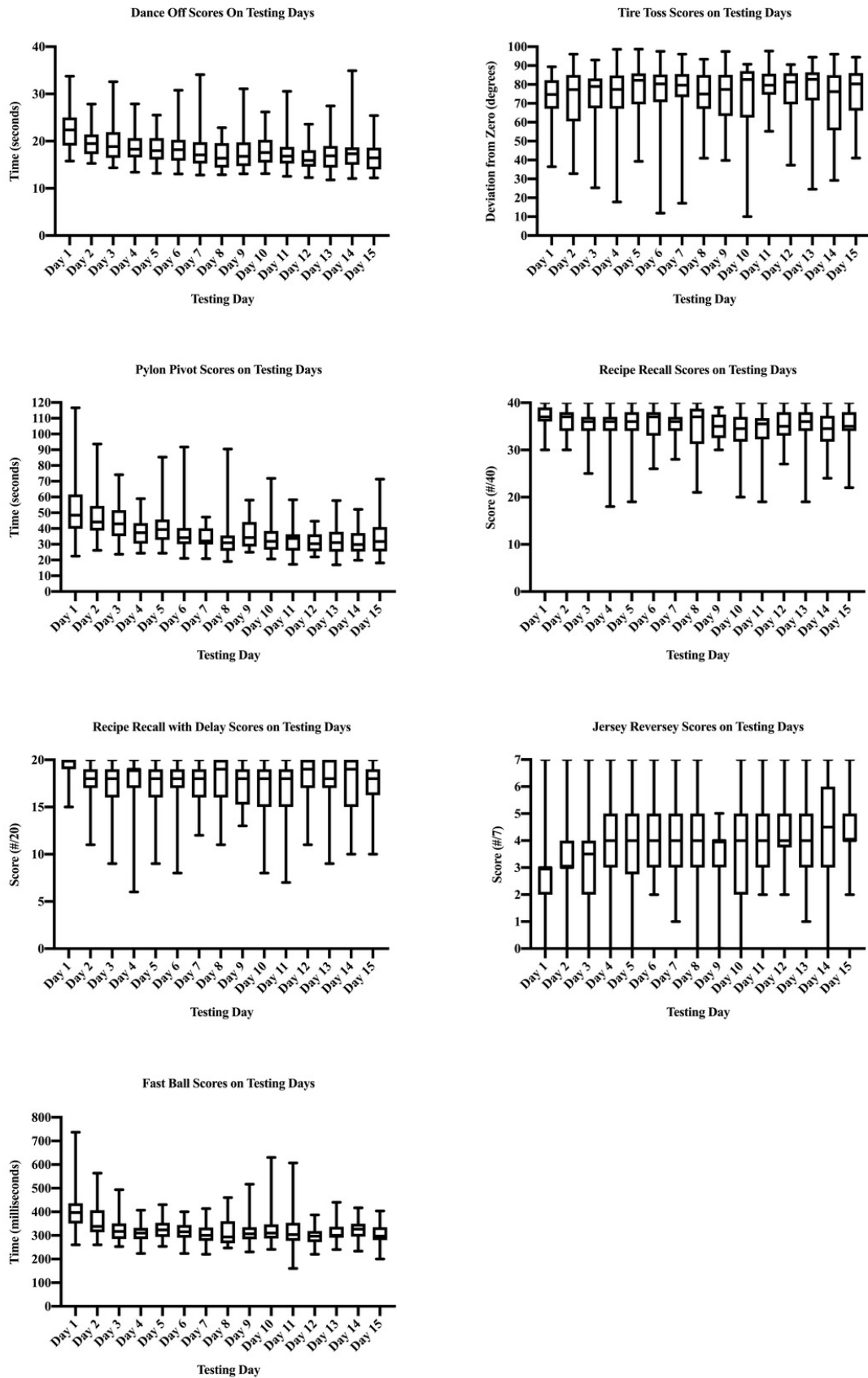


Figure 2. Scores, per subtest, across Testing Days 1-15. Whiskers identify minimum and maximum values.

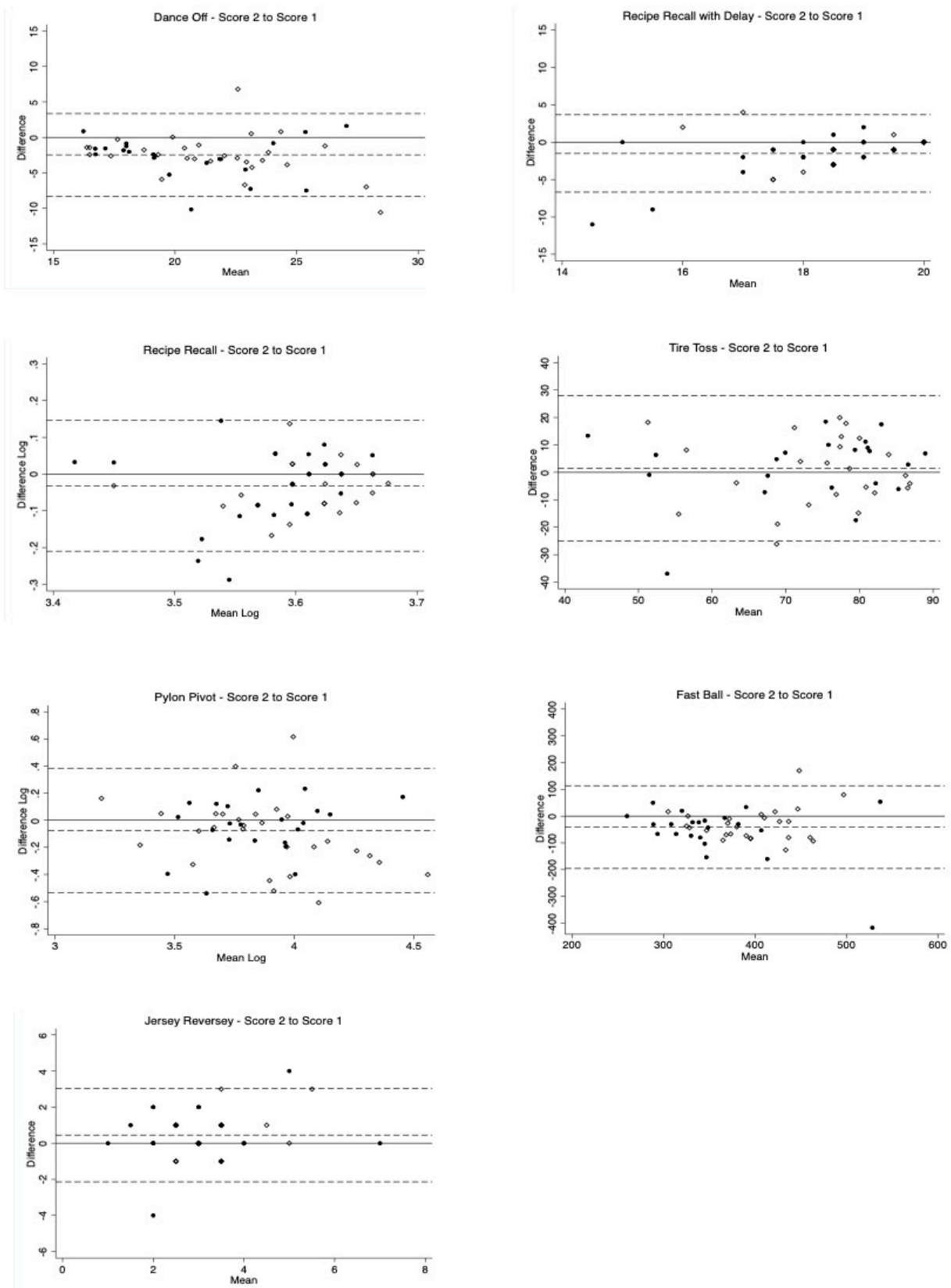


Figure 3. Bland Altman 95% Limits of Agreement plots for score 1 to score 2. Scores were adjusted for start time, indicating each individual's first and second testing administrations. Males are represented by black circles. Females are represented by white diamonds. Where assumptions were violated, scores were logarithmically transformed and logarithmic values are presented (Recipe Recall, Pylon Pivot).

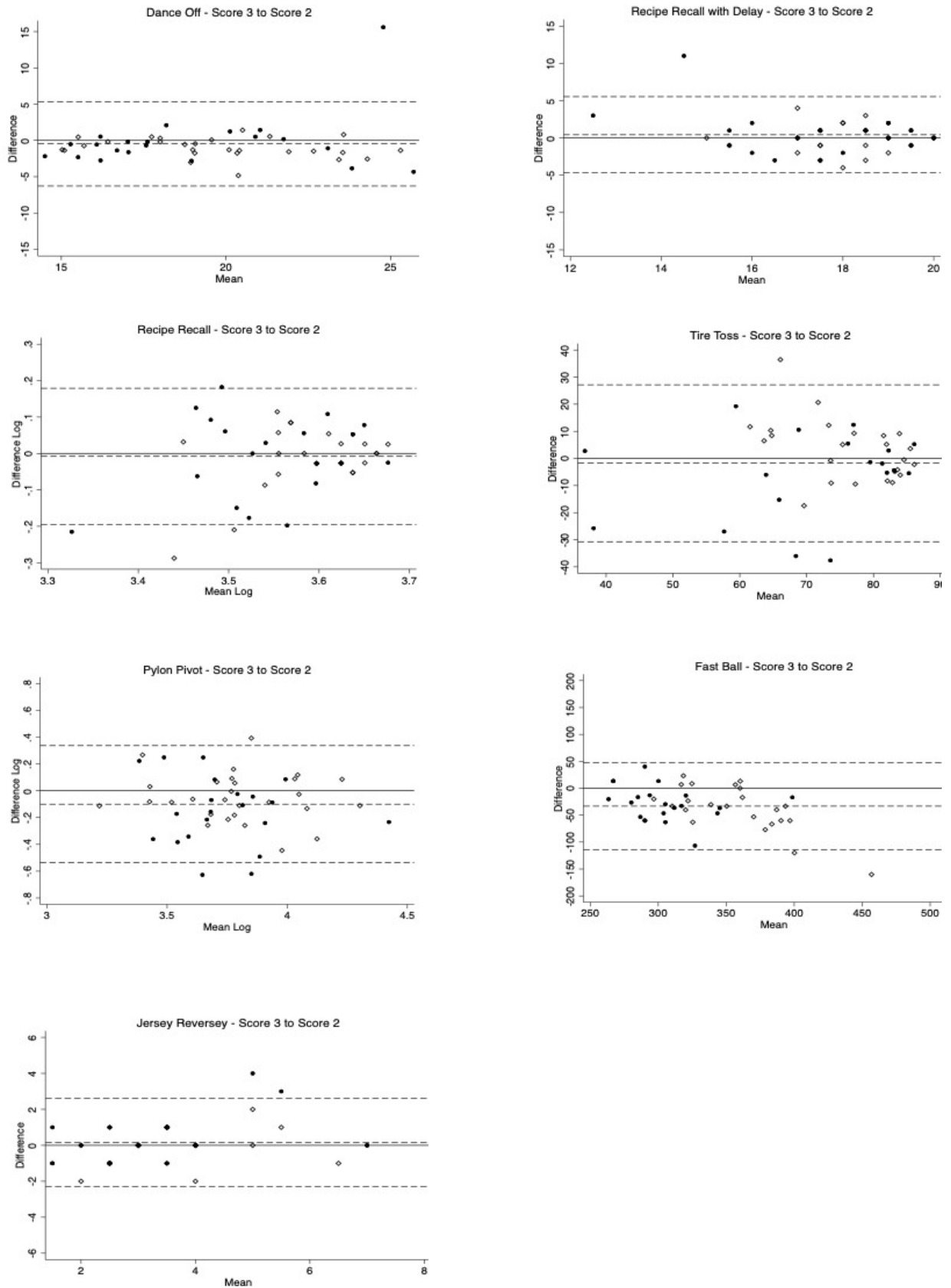


Figure 4. Bland Altman 95% Limits of Agreement plots for score 2 to score 3. Scores were adjusted for start time, indicating each individual's second and third testing administrations. Males are represented by black circles. Females are represented by white diamonds. Where assumptions were violated, scores were logarithmically transformed and logarithmic values are presented (Recipe Recall, Pylon Pivot).

Table 2. Bland Altman 95% Limits of Agreement

Subtest	Scores	Participant (n)	Range of Scores	Mean Difference (95% Confidence Interval)	Limits of Agreement
Dance Off	1 – 2	55	16.23 to 28.45	-2.46 (CI -3.25 to -1.68)	-8.38 to 3.35
	2 – 3	53	14.50 to 25.70	-0.47 (CI -1.27 to 0.33)	-6.28 to 5.34
Jersey Reversey	1 – 2	53	1.00 to 7.00	0.45 (CI 0.10 to 0.81)	-2.12 to 3.04
	2 – 3	51	1.50 to 7.00	0.16 (CI -0.19 to 0.50)	-2.29 to 2.60
Recipe Recall with Delay	1 – 2	53	14.50 to 20.00	-1.51 (CI -2.22 to -0.80)	-6.69 to 3.67
	2 – 3	51	12.50 to 20.00	0.43 (CI -0.29 to 1.15)	-4.68 to 5.54
Tire Toss	1 – 2	54	32.34 to 94.63	1.48 (CI -2.13 to 5.09)	-24.97 to 27.92
	2 – 3	51	22.68 to 92.36	-1.78 (CI -5.85 to 2.29)	-30.72 to 27.16
Fast Ball	1 – 2	55	260.21 to 537.01	-41.11 (CI -62.02 to -20.20)	-195.81 to 113.59
	2 – 3	53	263.45 to 520.44	-33.23 (CI -44.40 to -22.07)	-114.26 to 47.78
Recipe Recall	1 – 2	54	30.50 to 39.50	0.97 (0.95 to 0.99)*	0.81 to 1.16**
	2 – 3	51	28.00 to 39.50	0.99 (0.97 to 1.02)*	0.82 to 1.20**
Pylon Pivot	1 – 2	55	24.46 to 97.35	0.93 (0.87 to 0.99)*	0.59 to 1.46**
	2 – 3	54	25.02 to 83.93	0.90 (0.85 to 0.96)*	0.58 to 1.40**

* Geometric mean with 95% confidence interval presented when Bland and Altman plots did not meet the assumptions that the differences are normally distributed

** Limits of agreement are ratios for anti-log values presented (e.g., 0.81 to 1.16 interpreted as 95% of the time score 2 is 19% lower to 16% higher than score 1)

† Subtest Scoring:

a. Dance Off scores are calculated in seconds. Lower scores indicate better performance.

b. Jersey Reversey scores are calculated as number correct out of a possible 7. Higher scores indicate better performance.

c. Recipe Recall with Delay scores are calculated as number correct out of a possible 20. Higher scores indicate better performance.

d. Tire Toss scores are calculated as deviation from centre, in degrees, with a maximum score of 100. Higher scores indicate better performance.

e. Pylon Pivot scores are calculated in seconds. Lower scores indicate better performance.

f. Recipe Recall scores are calculated as number correct out of a possible 40. Higher scores indicate better performance.

g. Fast Ball scores are calculated in milliseconds. Lower scores indicate better performance.

gested that mean performance on timed subtests stabilized at day 3 for Dance Off and Fast Ball and day 4 for Pylon Pivot. This is similar to results described by Hinton-Bayre and colleagues, who noted that practice effects levelled off after a second baseline assessment.¹⁸ Further, repeat testing may have resulted in motor learning on balance tests (Pylon Pivot).¹⁹

Bland-Altman limits of agreement were wide, and where ICCs were calculated, results indicated poor-to-moderate reliability between scores 1 and 2 (Figure 3) and scores 2 and 3 (Figure 4).¹⁷ Scores for some subtests violated statistical assumptions, limiting the ability to quantify test-retest reliability. However, on the whole, the reliability of the HIEQ test battery across the first several administrations was limited. These results are similar to studies of other computerized test batteries used for concussion assessment, many of which have also demonstrated limited reliability of test scores.^{4,5} Together with improvements in mean performance on some tests, the results also suggest that the first one or two assessments might need to be dis-

regarded when establishing a multiple testing baseline.¹⁹ The findings of this study would also suggest that the first completion of the test should not be included in the calculation of a multiple baseline score given the potential for a learning effect.

To examine the reliability of multiple baselines, ICCs for each subtest were examined in two score bands (i.e., scores 2 through 6 and 7 through 11). The score bands were selected to reduce potential learning effects (score 1 to score 2) and to capture as many participant scores as possible. Where ICCs were calculated, results indicated poor-to-moderate reliability for both score bands. Thus, in this group administration setting, scores on the HIEQ showed only modest reliability over time.

Some authors have reported substantial individual variability on the Standardized Assessment of Concussion and the modified Balance Error Scoring Scale subtests of the SCAT5, with moderate reliability in a two-week test-retest reliability study.^{15,20} These findings suggest that multiple baselines may better capture the variability of performance

Table 3. Intraclass Correlations for Test-Retest Reliability – Score 1 to Score 2 and Score 2 to Score 3

Subtest	Scores	ICC	95% Confidence Interval		F-Test			
			Lower	Upper	df1	df2	F	p-value
Dance Off	1-2							
	2-3							
Recipe Recall	1-2	0.18	0-.06	0.42	53	53	1.50	0.073
	2-3	0.44	0.18	0.64	50	50	2.54	0.001
Pylon Pivot	1-2							
	2-3	0.64	0.40	0.79	53	53	5.30	<0.001
Jersey Reversey	1-2	0.52	0.29	0.70	52	52	3.42	<0.001
	2-3	0.68	0.51	0.82	50	50	5.34	<0.001
Recipe Recall with Delay	1-2							
	2-3							
Tire Toss	1-2	0.61	0.42	0.76	53	53	4.16	<0.001
	2-3	0.57	0.35	0.73	50	50	3.61	<0.001
Fast Ball	1-2							
	2-3	0.66	0.21	0.84	52	52	7.50	<0.001

Table 4. Intraclass Correlations for Reliability – Score 2 to Score 6 and Score 7 to Score 11

Subtest	Scores	ICC	95% Confidence Interval		F-Test			
			Lower	Upper	Df1	Df2	F	p-value
Dance Off	2 - 6							
	7 - 11							
Recipe Recall	2 - 6	0.34	0.21	.50	44	176	3.71	<0.0001
	7 - 11	0.51	0.34	.68	29	116	6.06	<0.001
Pylon Pivot	2 - 6							
	7 - 11	0.49	0.33	.66	31	124	5.85	<0.001
Jersey Reversey	2 - 6	0.55	0.41	.68	44	176	7.26	<0.001
	7 - 11	0.33	0.17	.52	29	116	3.41	<0.001
Recipe Recall with Delay	2 - 6							
	7 - 11	0.49	0.32	.67	28	112	5.71	<0.001
Tire Toss	2 - 6	0.53	0.40	.67	43	172	6.67	<0.001
	7 - 11							
Fast Ball	2 - 6							
	7 - 11	0.40	0.23	.61	25	100	4.53	<0.001

in an uninjured athlete, and this may eventually facilitate better detection and management should an athlete sustain a concussion. Further investigation into normative ranges for each individual, as part of a comprehensive multiple baseline assessment, may help to identify expected variability across differing domains (i.e. cognitive, balance, visual function) and may help to inform the optimal number of baseline testing days required for the different domains.

SPECIFIC SUBTEST CONSIDERATIONS

Scores on the memory subtests (Recipe Recall and Recipe Recall with Delay) were characterized by a ceiling effect, with many participants reaching the maximum score (40/40

or 20/20). This is similar to a ceiling effect identified when a five-word recall list was used for the memory components on the SCAT3 and on verbal memory subtests on the IMPACT test.^{21,22}

An interference effect also may have affected scores on memory subtests as participants completed the test over 15 testing days, with words presented on previous testing days interfering and affecting memory for current words, and thus resulting in a decrease in performance. Although these subtests were developed as a more functional task (grocery shopping list) and may be more applicable to daily life settings, a larger or more varied word bank may help to prevent interference effects.

LIMITATIONS

The study had several limitations. School barriers, such as school closure, personal activity days, exams, and lost internet connections hampered completion of the HIEQ on some testing days. The test was performed in a classroom setting; therefore, the participants may have been distracted or given less effort than they would have in a distraction-free setting without others present. This may have led to increased variability in test scores, although the use of headphones would reduce distraction. Evaluation of reliability of repeat testing in visual and auditory distraction free individual testing environments is warranted and is in keeping with current testing recommendations. Participants were asked to complete the test for 15 consecutive school days; the number of consecutive days may have decreased engagement and also contributed to variability in test scores, although reliability was not noticeably different across the two testing bands. In addition, some participants elected to complete only certain subtests or some subtests repeatedly, which may have influenced their exposure to certain subtests, and thus the reliability of their performance. Randomizing the order of the subtests may increase engagement with the HIEQ test battery. Further, two schools in the study included high-performance student-athletes. The performance of adolescents attending high-performance schools may differ from that of adolescents attending a typical high school. As participation in the study was voluntary and recruitment happened through specific schools and classes a selection bias may have occurred where the students who performed at a higher level or were more likely to perform well on the tests may have chosen to participate. The study was not adequately powered to permit evaluation of differences in performance between schools.

CONCLUSIONS

While the HIEQ appears to be feasible in high school students, test-retest reliability is poor-to-moderate when the tests are administered in a group setting. Fifteen days of repeated baseline testing may be burdensome to adolescents. Further research evaluating five-day multiple baseline testing of individuals in visual and auditory distraction free individual testing settings may provide greater insight into the reliability of the HIEQ in uninjured adolescents athletes.

DECLARATION OF INTEREST

Kathryn Schneider is a physiotherapist consultant at Evidence Sport and Spinal Therapy and at University of Calgary Sport Medicine Centre.

FUNDING

We acknowledge funding through Highmark Innovations Inc. (Toronto, Canada) for the support of this project and the studentship of HAS.

CONFLICTS OF INTEREST

There are no conflicts of interest to declare.

APPROVAL

This study was approved by the University of Calgary Conjoint Health Research Ethics Board for human subject research.

ACKNOWLEDGEMENTS

The Sport Injury Prevention Research Centre is one of the International Research Centres for Prevention of Injury and Protection of Athlete Health supported by the International Olympic Committee. We acknowledge funding from Highmark Innovations Inc. (Toronto, Canada) for the support of this project. We acknowledge Vineetha K. Waryiar for her contributions to the project. Keith Yeates holds the Ronald and Irene Ware Chair in Pediatric Brain Injury, funded by the Alberta Children's Hospital Foundation. Carolyn Emery holds a Canada Research Chair (Tier 1) in Concussion. We acknowledge the support of the students, teachers, and schools who participated in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to the conceptualization and design of the study. HAS, CRVR, and KJS led the protocol development for data collection and HAS, CRVR, and RFG conducted all data collection and data cleaning. HAS conducted all data analysis with mentorship from all team members in particular AMB, KOY, and CAE. The paper was initially drafted by HAS. HAS and KJS take responsibility for the integrity of the data and the accuracy of the data analysis and manuscript. KJS takes responsibility for funding acquisition. All authors critically reviewed and edited the manuscript before submission.

Submitted: September 13, 2021 CDT, Accepted: December 08, 2021 CDT



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-NC-4.0). View this license's legal deed at <https://creativecommons.org/licenses/by-nc/4.0> and legal code at <https://creativecommons.org/licenses/by-nc/4.0/legalcode> for more information.

REFERENCES

1. Black A, McCallum J, Eliason PE, Hagel BE, Emery CA. The burden of sport-related concussion in high school students: lifetime prevalence, 1-year incidence and persistent symptoms. *Clin J Sport Med*. 2020;30(3):e89. doi:10.1097/JSM.0000000000000844
2. McCrory P, Meeuwisse W, Dvořák J, et al. Consensus Statement on Concussion in Sport—the 5th International Conference on Concussion in Sport held in Berlin, October 2016. *Br J Sports Med*. 2017;51(11):838. doi:10.1136/bjsports-2017-097699
3. Feddermann-Demont N, Echemendia RJ, Schneider KJ, 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. doi:10.1136/bjsports-2016-097403
4. Farnsworth JL, Dargo L, Ragan BG, Kang M. Reliability of computerized neurocognitive tests for concussion assessment: A meta-analysis. *J Athl Train*. 2017;52(9):826-833. doi:10.4085/1062-6050-52.6.03
5. Kontos A, Sufrinko A, Womble M, Kegel N. Neuropsychological assessment following concussion: An evidence-based review of the role of neuropsychological assessment pre- and post-concussion. *Curr Pain Headache Rep*. 2016;20(6):1-7. doi:10.1007/s11916-016-0571-y
6. Bruce J, Echemendia R, Tangeman L, et al. Two baselines are better than one: Improving the reliability of computerized testing in sports neuropsychology. *Appl Neuropsychol Adult*. 2016;23(5):336-342. doi:10.1080/23279095.2015.1064002
7. Echemendia RJ, Meeuwisse W, McCrory P, et al. The Sport Concussion Assessment Tool 5th Edition (SCAT5): Background and rationale. *Br J Sports Med*. 2017;51(11):848. doi:10.1136/bjsports-2017-097506
8. McCrea M. Standardized mental status testing on the sideline after sport-related concussion. *J Athl Train*. 2001;36(3):274-279.
9. Partington JE, Leiter RG. Partington's Pathways Test. *Psychol Serv Cent J*. 1949;1:11-20.
10. Galetta KM, Liu M, Leong DF, Ventura RE, Galetta SL, Balcer LJ. The King-Devick Test of rapid number naming for concussion detection: Meta-analysis and systematic review of the literature. *Concussion*. 2016;1(2):cnc.15.8. doi:10.2217/cnc.15.8
11. King-Devick Technologies, Inc. Accessed April 9, 2021. <https://kingdevicktest.com/>
12. Highmark Interactive. EQ Science . Published 2020. Accessed August 5, 2020. <https://www.highmark.tech/science/>
13. Highmark Interactive. EQ Active. Published 2020. Accessed August 5, 2020. <https://www.highmark.tech/eqactive/>
14. Gokula KA. Romberg's Test. *J Postgrad Med*. 2003;49:169-172.
15. Bell DR, Guskiewicz KM, Clark MA, Padua DA. Systematic review of the Balance Error Scoring System. *Sports Health*. 2011;3(3):287-295. doi:10.1177/1941738111403122
16. Khanna NK, Baumgartner K, Labella CR. Balance Error Scoring System performance in children and adolescents with no history of concussion. *Sports Health*. 2015;7(4):341-345. doi:10.1177/1941738115571508
17. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
18. Hinton-Bayre AD, Geffen GM, Geffen LB, McFarland KA, Frijs P. Concussion in contact sports: Reliable change indices of impairment and recovery. *J Clin Exp Neuropsychol*. 1999;21(1):70-86. doi:10.1076/jcen.21.1.70.945
19. Manaseer TS, Whittaker JL, Isaac C, Schneider K, Roberts MR, Gross DP. The reliability of clinical balance tests under single-task and dual-task testing paradigms in uninjured active youth and young adults. *Int J Sports Phys Ther*. 2020;15(4):487. doi:10.26603/ijsp20200487
20. Luoto T, Parkkari J, Seppänen A, Palola V, Tuominen M, Iverson GL. Reliability of the Sport Concussion Assessment Tool 5 baseline testing: A 2-week test-retest study. 2019;(March):2-3. doi:10.26226/MORRESSIER.5C7E3E2129D813000CB41BA0
21. Snedden TR, Brooks MA, Hetzel S, McGuine T. Normative values of the Sport Concussion Assessment Tool 3 (SCAT3) in high school athletes. *Clin J Sport Med*. 2017;27(5):462-467. doi:10.1097/JSM.0000000000000389
22. Gaudet CE, Konin J, Faust D. Immediate Post-concussion and Cognitive Testing: Ceiling effects, reliability, and implications for interpretation. *Arch Clin Neuropsychol*. 2021;36(4):561-569. doi:10.1093/A RCLIN/ACAA074