*Journal of Athletic Training* 2011:46(5):505–513
© by the National Athletic Trainers' Association, Inc
www.nata.org/jat

# Reliability of Fitness Tests Using Methods and Time Periods Common in Sport and Occupational Management

**Bryan D. Burnstein, MS, ATC, CSCS, NASM-PES*; Russell J. Steele, PhD†; Ian Shrier, MD, PhD‡**

*Cirque du Soleil, Las Vegas, NV; †Department of Mathematics and Statistics and ‡Centre for Clinical Epidemiology and Community Studies, Jewish General Hospital, McGill University, Montreal, QC, Canada

***Context:*** Fitness testing is used frequently in many areas of physical activity, but the reliability of these measurements under real-world, practical conditions is unknown.

***Objective:*** To evaluate the reliability of specific fitness tests using the methods and time periods used in the context of real-world sport and occupational management.

***Design:*** Cohort study.

***Setting:*** Eighteen different Cirque du Soleil shows.

***Patients or Other Participants:*** Cirque du Soleil physical performers who completed 4 consecutive tests (6-month intervals) and were free of injury or illness at each session ($n = 238$ of 701 physical performers).

***Intervention(s):*** Performers completed 6 fitness tests on each assessment date: dynamic balance, Harvard step test, handgrip, vertical jump, pull-ups, and 60-second jump test.

***Main Outcome Measure(s):*** We calculated the intraclass coefficient (ICC) and limits of agreement between baseline and each time point and the ICC over all 4 time points combined.

***Results:*** Reliability was acceptable ($ICC > 0.6$) over an 18-month time period for all pairwise comparisons and all time points together for the handgrip, vertical jump, and pull-up assessments. The Harvard step test and 60-second jump test had poor reliability ($ICC < 0.6$) between baseline and other time points. When we excluded the baseline data and calculated the ICC for 6-month, 12-month, and 18-month time points, both the Harvard step test and 60-second jump test demonstrated acceptable reliability. Dynamic balance was unreliable in all contexts. Limit-of-agreement analysis demonstrated considerable intraindividual variability for some tests and a learning effect by administrators on others.

***Conclusions:*** Five of the 6 tests in this battery had acceptable reliability over an 18-month time frame, but the values for certain individuals may vary considerably from time to time for some tests. Specific tests may require a learning period for administrators.

***Key Words:*** fitness assessment, dynamic balance, Harvard step test, handgrip, vertical jump, pull-ups, 60-second jump test

**Key Points**

- For fitness testing under real-world sport and occupational settings and over time, the modified Harvard step, handgrip, vertical jump, pull-up, and 60-second jump tests were reliable; the dynamic balance test was not. However, the Harvard step and 60-second jump tests demonstrated learning effects.
- Although these tests may be useful in identifying differences among individuals, limits-of-agreement analysis indicated that the tests are restricted in their ability to detect conditioning changes in individuals over time.

Fitness testing is a commonly used management tool in a wide variety of organizations that involve physical activity, including sports teams, the military, and police and fire departments. It provides information that can help to assess the ability to perform a required job or sport tasks,[1,2] track conditioning or deconditioning over time,[3] evaluate the effectiveness of strength and conditioning interventions,[4,5] increase participant motivation,[4] and identify strengths and weaknesses so that fitness or injury-prevention programs can be tailored appropriately.[2] Fitness testing consists of measuring different aspects of physical health and performance with objective assessments. The most common areas of interest include flexibility, strength (absolute, maximal, or relative), agility, power (anaerobic capacity), and endurance (aerobic capacity).[6] All or only some of these aspects may be included in a fitness assessment, depending on which specific objective is the goal. For example, police departments might be interested in focusing on levels of endurance and upper body strength (eg, for foot chases),[2] whereas basketball teams would probably be interested in endurance, agility, and power (jump height).[7]

After choosing the specific types of tests that are required, one needs to select a specific test for each aspect of fitness. A test should be considered for use only if the necessary equipment is available, the testing environment is appropriate, and the test meets the goals of the assessment and is appropriate for the age and physical limitations of the participants. Of the tests being considered, the most common criteria used to select a specific assessment are its validity and reliability. In general, test-retest reliability is measured in a laboratory setting under very controlled conditions,[8–14] but the tests are ultimately used under less than ideal conditions (eg, by many administrators, testing at different times of day). In addition, most reliability testing is conducted over a short period of time, but the reliability properties may be very different over the months to years in which they are routinely used in the field.[15] Therefore, the objective of our study was to assess some typical fitness tests for reliability within the environment and using the methods and time frame that will be used in the field.

To address this question, we used relevant data obtained from Cirque du Soleil (CdS) during its normal business practices. Since 2007, all full-time and temporary CdS artists in 18 different shows have participated in physical capacity assessments (PCAs) every 6 months that included tests similar to those conducted in elite athletes. These tests address proprioception, aerobic fitness, general upper body strength, lower body power, upper body strength and endurance, and anaerobic capacity. Furthermore, testing in the CdS context is similar to that performed by individual teams within a sport conducting their own tests, with clinicians rather than trained researchers administering the tests.

## METHODS

### Participants

Cirque du Soleil includes both athletic and artistic performers, and all were tested. We excluded all clowns, characters, and musicians whose activities do not require a high level of physical fitness; the current analysis was limited to physical performers whose primary role consisted of athletic, acrobatic, or dance maneuvers that often involve sudden compression or distraction loads similar to those found in sports and other physically demanding professions. In order to assess test-retest reliability over a long period of time, we chose to analyze only data from physical performers who completed testing at 4 consecutive time points (6-month intervals) and were free of injury or illness at each session.

After consultation, our research ethics committee determined that because this project used historical data from the records of a private company to assess the performance of the company and its employees and the data were not gathered for research purposes, the study fell under a category of quality assurance that did not require formal research ethics approval.

### General Procedures

The PCAs were administered by each show's Performance Medicine Department personnel (certified athletic trainers, athletic therapists, and physical therapists) and coaching staff at the location of the show. Approximately 50% of the CdS shows remain in one location (eg, Las Vegas, NV; Orlando, FL), and the remaining (touring) shows travel the world. Each test administrator viewed a training video, was given a detailed handbook, and underwent the series of tests during a summit. When staffing changes occurred, the procedures were reviewed in a conference call and all questions addressed.

Staff provided verbal motivation at each station. Whenever possible, the same staff person at each show worked the same test station each time PCAs were administered. Staff also documented the perceived effort of each performer as *appropriate* or *inappropriate submaximal* (artists who performed the tests while injured or ill were considered appropriate submaximal but excluded from this analysis).

Testing was conducted on a regular workday, either in the early afternoon or after a show to avoid interfering with the performer's ability to safely complete his or her routine. Each individual show kept this schedule consistent, although some variation existed because of the performing arts schedule. Testing was also scheduled so that it did not occur within the first 2 weeks after an extended period of time off or, in the case of touring shows, a new city. On test days, all performers were instructed to wear comfortable, athletic-type clothing and be well hydrated and fed. Cirque du Soleil used a multistation design in which performers underwent 1 test at each station. In general, performers had 1 to 5 minutes of rest between stations, depending on how many performers were being tested and the duration of each test. For the proprioception assessment (dynamic balance), performers were allowed 1 practice trial per leg, with a maximum of 15 seconds per leg. Practice trials were not offered for any other test.

### Tests

The tests were chosen by CdS to provide a snapshot view of general overall fitness and are not considered to be act or sport specific. These included tests for proprioception, aerobic fitness, general upper body strength, lower body power, upper body strength and endurance, and anaerobic capacity. The specific tests were selected from the published literature and were considered commonly used in sport. Each test is described in this section.

Proprioception is considered a key element for optimal human performance and injury prevention. Cirque du Soleil opted for a dynamic balance test with eyes closed[14] because it requires minimal equipment (a specialized foam pad) and approximately 5 seconds to complete, and it has been used in injury-prevention research.[16] The performer stood on a high-density balance pad (Balance Pad Elite; Airex Specialty Foams, Aargau, Switzerland) with the hands on the waist and wearing no shoes. The test began when the performer lifted 1 leg, ensuring that it did not touch the opposite leg, and closed his or her eyes. The objective was to stand for as long as possible. The test was considered completed if any of the following occurred: the performer removed 1 or both hands from the hips, the non-support leg touched the foam or floor, the weight-bearing foot moved from its original position, the eyes were opened, or 5 seconds of swaying occurred. Each performer was allowed (but not required to use) 1 practice trial lasting up to 15 seconds per leg before beginning the test. The maximum time for this test is 180 seconds, and the better performance of 2 trials was recorded.

Among the numerous tests available for aerobic fitness, CdS selected a modified Harvard step test[8] because it takes little space, time, and equipment. This test is based on the premise that performers with higher fitness levels have smaller increases in heart rate with stepping up and down a 44-cm step at a cadence of 100 beats per minute (up on 1 and 2, down on 3 and 4) for 5 minutes, as well as faster recovery times. The results are strongly correlated with maximal oxygen consumption.[8] In order to avoid soreness in performers who had to participate in shows the same night, CdS used a modified step height (40 cm for performers taller than 137 cm and 33 cm for other performers), and performers were allowed to switch the step-up leg throughout the test. After 5 minutes, the performer immediately sat down and remained as still and quiet as possible. The overall score is 30000 divided by the sum of the heart rates obtained by heart rate monitors (several models; Polar Electro USA, Lake Success, NY) at 1 minute, 2 minutes, and 3 minutes after testing. Data from performers who were not able to complete the full 5 minutes were considered missing. In order to motivate the performers, each box was wide enough to accommodate 2 artists at the same time (however, at some test periods, an odd number of artists were present; at others, an artist's schedule required special accommodation), and popular music that was time coded to 100 beats per minute was played.

General upper body strength was tested with the frequently used handgrip test.[12] Handgrip is important for activities such as throwing and catching, which are necessary to many circus acts. Grip strength is also correlated with other strength measures, such as elbow flexion, knee extension, trunk flexion, and trunk extension.[17] A hand dynamometer (Baseline; Fabrication Enterprises, Inc, White Plains, NY) was used to measure grip strength. The performer was positioned so that the heels, buttocks, shoulders, and back of the head were flat against the wall, the shoulder was adducted and in neutral rotation, the elbow was flexed to 90°, and the lower arm and wrist were in neutral position.[11] Unlike the rest of the body, the elbow was not allowed to touch the wall. The dynamometer was sized to the individual with its spine parallel to the performer's thumb. If the performer's longest finger was shorter than the dynamometer's poles, the handle was then set to position 2 (1-7/8 in [4.76 cm] from the dynamometer spine), and if the performer's longest finger was longer than the poles, position 3 (2-3/8 in [6.03 cm] from the dynamometer spine) was used. The performer gripped the dynamometer (with or without chalk, as desired) while keeping the wrist neutral and squeezed for 3 to 5 seconds. Any pumping of the dynamometer was considered a failed trial and was redone because it seemed to cause a false high reading. The maximum score was 90.7 kg (200 lb), and the better of 2 trials was recorded.

One of the more accepted tests for lower extremity power is the vertical jump test.[10] Cirque du Soleil used the Vertec (Jump USA; Sports Imports, Columbus, OH) to measure countermovement jump height.[10] The test is performed by having the performer stand with both feet flat on the floor, drop the arms, and then flex at the hip, knee, and ankle before exploding upward at takeoff with the objective of touching the highest possible vane. The use of a drop-step technique was encouraged. This test began as did the traditional countermovement jump, but some lower extremity motion was allowed before the jump occurred. The performer was allowed to step back with one foot and then return the same foot to the initial starting position before jumping. His or her standing reach height while barefoot was recorded, allowing for trunk side flexion, which also occurs during the test. The jump height is the difference between the height of the highest vane that moved and the performer's reach height. The better of 2 trials was recorded.

Cirque du Soleil measured upper body muscular strength and endurance with the pull-up test[18] because it takes little space, equipment, and time. In addition, many circus acts demand this type of movement during shows. Performers began the test from a full hang off the pull-up bar, with the palms facing away from them (ie, overhand grip) roughly shoulder-width apart (with or without chalk as desired). For a successful pull-up, the chin cleared the bar; attempts associated with body swinging, absence of full arm extension when returning to the starting position, or lifting the chin (neck extension) were excluded. A score of 0.5 on the final attempt was recorded if the elbow joint reached 90° of flexion.

Hoffman and Kang[19] validated certain field tests for anaerobic capacity. Cirque du Soleil chose to use a simple 60-second jump test reported on the Internet (not validated to our knowledge) because it is more sport specific to performers' tasks than cycle ergometer tests and involves less equipment.[20] For the jump test, the performer completed as many successful lateral hops back and forth as possible. Three parallel lines were marked on the floor with athletic tape, 30 cm apart. To start, the performer stood on the center line with the feet close together. The performer then jumped from line to line with both feet together (one cycle is defined as jumping to the outside of the left line, back to center, to the outside of the right line, and then back to center) nonstop during 1 minute using a countdown timer. The administrator recorded the number of successful cycles in 1 minute. The test was not stopped for incorrectly completed cycles, but they were not counted in the final score. If time ran out and the performer was in the middle of a successful cycle, an additional 0.5 was added to the final score.

## Data Analysis

We describe the demographics of our population using mean ± standard deviation for continuous variables and percentages for categoric variables. The data for the dynamic balance test and pull-ups were highly skewed and showed heteroscedasticity. These variables were therefore log transformed for all analyses (performers who received a score of 0 for pull-ups were assigned 0.5).

We used 3 methods to assess test-retest reliability. First, we present box plots for each time point to provide an overview of the differences between each performer's score at that time point and the average of all of his or her scores. To do this, we calculated each performer's overall mean for the 4 time points. Then, for each session, we subtracted this mean from the result for that session; a score of 0 meant the result for that session was equal to the mean of the 4 sessions for that individual. In order to provide the reader with more information, we then added the overall mean of all performers for that session to each score so that the box plot was centered at the value of the session mean. We calculated the intraclass coefficient (ICC) (2,1) for agreement[21] (with 95% confidence intervals using the bootstrap percentile method[22]) over all 4 time points combined. Because of an apparent learning effect for some tests between baseline and 6 months, we also calculated the ICCs and 95% confidence intervals across the combined time points of 6 months, 12 months, and 18 months. For transparency reasons only, we provide the ICC (without the 95% confidence interval) for each variable for all pairwise comparisons: baseline versus 6 months, baseline versus 12 months, baseline versus 18 months, 6 months versus 12 months, 6 months versus 18 months, and 12 months versus 18 months. Because our study's measurements took place over a long period of time under less stringent conditions, we accepted a value of 0.6 as the minimum for acceptable reliability,[15] which is lower than the acceptable value for reliability tested under more strict conditions.

Finally, we examined the 95% limits of agreement.[23] In brief, the limits of agreement measure the magnitude of the variability (standard deviation) of the difference between scores for individuals at different testing sessions. This information is extremely helpful in determining how much one would expect the value for a person to vary by chance with every test. Plotting these differences against the average of the scores for the individual allows us to easily see whether the magnitude of the difference depends on the absolute value of the score achieved; this information is not available from the ICC calculations. The average is used for this plot because no single value is "more correct" than any other value, and the average represents the best estimate of the true value for that person. The main underlying assumption of this analysis is that the variability for each testing session is approximately the same, which is reasonable in the context of measuring the same participants on the same test at different time periods.

Because our preliminary data suggested that the baseline testing was not reliable, we restricted the limits-of-agreement analysis to the comparison of 18-month and 6-month data; we conducted sensitivity analysis for 18-month to 12-month data as well. We also performed a sensitivity analysis on the data of only the participants who demonstrated appropriate effort and included analyses in which the data were or were not log transformed. We used open-code statistical software (2007 version; The R Project for Statistical Computing, Wirtschafts Universitat, Vienna, Austria) for all analyses.

## RESULTS

Of the 809 performers tested at baseline, 108 were excluded because they were not physical performers (eg, clowns, musicians, characters). Of the 701 physical performers, 463 were excluded because they either missed a testing session for any reason or could not give an appropriate maximal effort because they were considered ill or injured during 1 of the 4 testing periods. The baseline demographic data for the 238 included healthy physical performers and 463 excluded physical performers are shown in Table 1.

Box plots showing the variability across trials for each performer are displayed in Figure 1. In addition, we plotted the overall mean for all performers across all 4 time points (dotted line) and provided the actual value of the overall mean ± standard deviation for each test at the top of the box plot. For most measures, the results in the box plots indicated no improvement over time, with the notable exceptions of lower values for balance tests at baseline and improvement over time for the 60-second jump test.

**Table 1. Demographic Information for Physical Performers**

| Variable | Included (n = 238) | Excluded[a] (n = 463) |
|---|---|---|
| Sex, n (%) | | |
| Men | 162 (68.1) | 290 (62.7) |
| Women | 76 (31.9) | 173 (37.3) |
| Location, n (%)[b] | | |
| Tour | 90 (37.8) | 222 (48.0) |
| Resident | 148 (62.2) | 235 (50.6) |
| International headquarters | 0 (0) | 7 (1.4) |
| Physical characteristics, mean ± SD | | |
| Height, m | 1.7 ± 0.1 | 1.7 ± 0.1 |
| Mass, kg | 64.8 ± 14.3 | 66.0 ± 17.7 |
| Body mass index, kg/m$^2$ | 22.8 ± 3.2 | 23.2 ± 4.3 |
| Age, y | 28.7 ± 6.4 | 29.6 ± 8.4 |

[a] Performers who did not complete all 4 rounds of testing because of current or recent injury.

[b] Cirque du Soleil shows are divided into touring shows, which change locations, and resident shows, which remain at permanent locations (eg, Las Vegas, NV). Some artists were tested at the international headquarters when they arrived at the company.

The ICC results for each pairwise comparison, for all 4 time points together, and for the 6-month through 18-month tests are shown in Table 2. Our results suggest acceptable test-retest reliability over long periods of time for each pairwise comparison and all time points together for the handgrip, vertical jump, and pull-up assessments. However, the pairwise ICC comparisons for the Harvard step test and the 60-second jump test indicated that the baseline measurements displayed poor reliability with other time points but that the reliability increased after that. When we excluded the baseline testing and calculated the ICC for 6 months through 18 months combined, both the Harvard step test and the 60-second jump test demonstrated acceptable reliability, with ICCs of 0.63 and 0.71, respectively. Finally, although the reliability of the dynamic balance test improved after the baseline testing, it never reached a level of acceptability for any pairwise comparison using the log-transformed data (Table 2) or the raw numbers (data not shown).

To assess whether poor motivation reduced the reliability coefficients, we conducted a subgroup analysis of only the performers who demonstrated an appropriate effort, as documented by the administrator at the time of the test (n = 215). However, the ICC was essentially unaffected (Table 2).

Based on the ICC values, we disregarded the baseline testing for the limits of agreement and instead compared the data collected at 6 months and 18 months. The limits of agreement for the left and right dynamic balance tests (using log-transformed data) are wide, with 2 SDs (left = 1.2, right = 1.4) representing approximately 1/3 to 1/2 of the overall mean value for the average across both tests (left = 3.2, right = 3.4) (Figure 2). Although the limits of agreement are smaller for the other tests, variability was still considerable for some performers tested repeatedly over time, suggesting that a particular person's score may vary substantially over time. The results were qualitatively similar when we compared 12 months with 18 months (data not shown).

Taken together, the ICC results indicated that these tests were reliable to distinguish among performers in a population, but the limits of agreement reflect wide variability for each performer when measured at different times.

## DISCUSSION

Test-retest studies often evaluate reliability in a well-controlled environment and over a short period of time. In our study, we tested reliability over an 18-month period and found it was acceptable (>0.6) for 5 of the 6 tests. The handgrip, vertical jump, and pull-ups had high levels of reproducibility throughout the 18 months. The reliability of the modified Harvard step test and the 60-second jump test increased substantially after the first testing session, indicating a learning effect. The reliability of the dynamic balance tests was poor for all comparisons. According to the limits of agreement, although reliability was acceptable for the different tests to assess population differences, considerable intraindividual variation remained from test to test, restricting the ability to detect clinically relevant changes within an individual based on a comparison of only 2 testing time points. To help clinicians make rational choices as to which tests to use, we discuss the reliability of each of the following tests in comparison with what is known about common alternatives. Unfortunately, most of the literature on limits

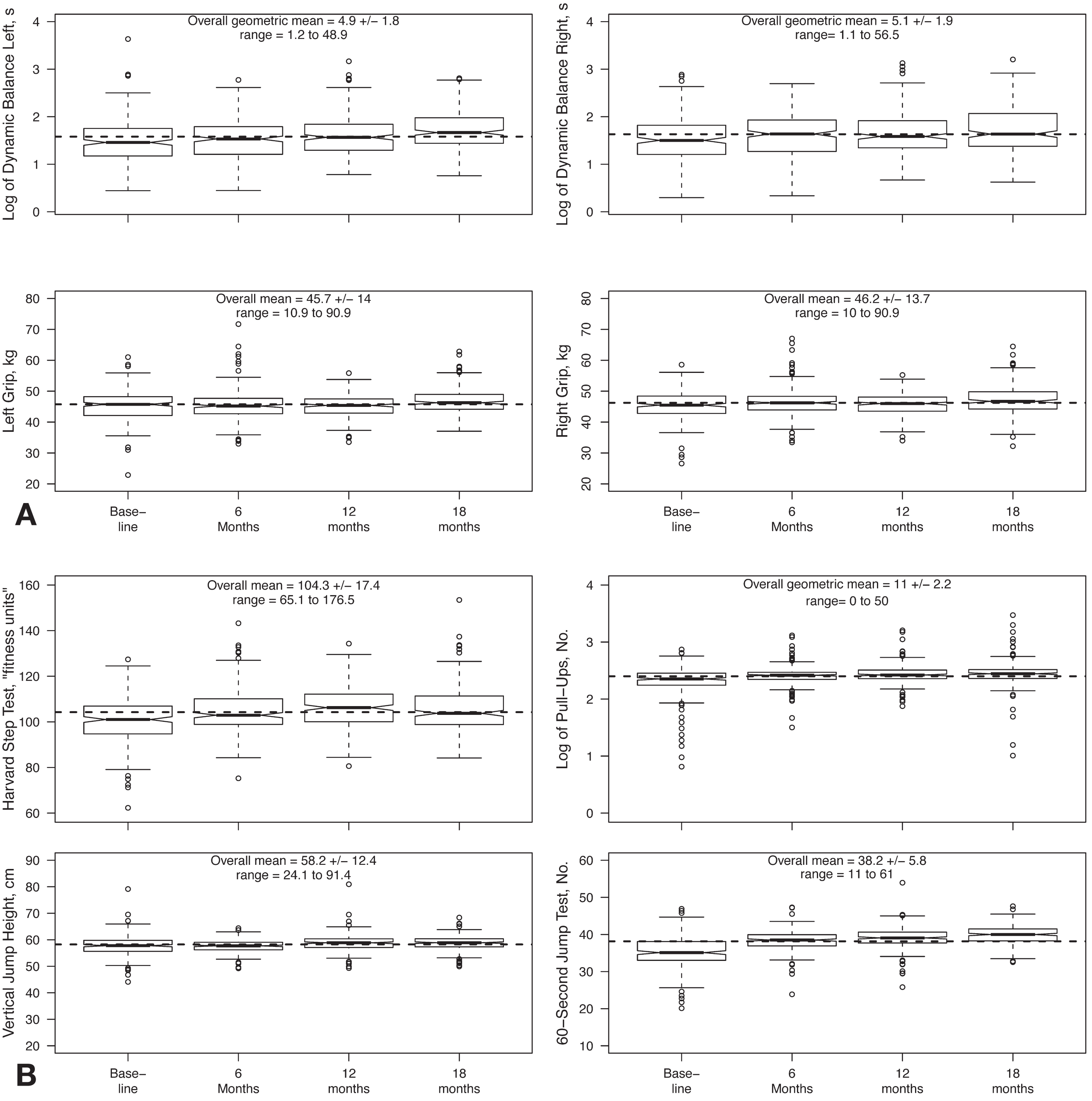

**Figure 1. Box plots for fitness tests across all time points. A, Tests applied to both right and left limbs. B, Other tests. Log-transformed data are provided for the balance test and pull-up test because the distributions were highly skewed. For any time point, the box indicates the 25th and 75th percentiles, and the dark bar represents the median. Notches around the median that do not overlap provide strong evidence for a difference in medians. The whiskers represent 1 interquartile region below and above the 25th and 75th percentiles, respectively, and the circles represent the outlying data points. The overall mean ± SD across all 4 time periods is indicated by the dotted line. We also provide the numeric value for the mean, SD, and range across all time points (or geometric mean [geometric SD] for log-transformed data) for each test.**

of agreement was restricted to a young pediatric population; therefore, our discussion for this type of analysis is limited to dynamic balance and handgrip tests. For all other tests, our limits-of-agreement analysis indicated that the variation of an individual's score over time meant that small changes in physical capacity could not be reliably detected.

We found high ICC values across all time-points for the pull-up (0.88), handgrip (left = 0.87, right = 0.85), and vertical jump tests (0.85). These high ICCs may be related to the ease of administering the test and the familiarity of the test for both performers and administrators. Despite its common use, we could not find any studies reporting the reliability for pull-ups. Our handgrip test results using the Baseline hand dynamometer were consistent with those of authors[11,24] who reported ICCs of 0.85 to 0.98 using the Jamar dynamometer (Sammons Preston Rolyan, Bolingbrook, IL). In addition to its high reliability

**Table 2. Intraclass Coefficients (ICCs) for the Comparison of Fitness Tests in Performers (n=238) Over Time**

| | Test | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Time Points | Dynamic Balance Left[a] | Dynamic Balance Right[a] | Handgrip Left | Handgrip Right | Vertical Jump | Pull-Ups[a] | Harvard Step | 60-Second Jump |
| Baseline versus 6 mo[b] | 0.23 | 0.06 | 0.86 | 0.86 | 0.92 | 0.90 | 0.58 | 0.49 |
| Baseline versus 12 mo | 0.25 | 0.29 | 0.89 | 0.89 | 0.88 | 0.87 | 0.56 | 0.39 |
| Baseline versus 18 mo | 0.23 | 0.24 | 0.85 | 0.81 | 0.89 | 0.84 | 0.46 | 0.34 |
| 6 mo versus 12 mo | 0.50 | 0.38 | 0.88 | 0.90 | 0.93 | 0.94 | 0.67 | 0.72 |
| 6 mo versus 18 mo | 0.44 | 0.35 | 0.84 | 0.83 | 0.94 | 0.91 | 0.60 | 0.66 |
| 12 mo versus 18 mo | 0.45 | 0.51 | 0.88 | 0.84 | 0.94 | 0.94 | 0.64 | 0.75 |
| Baseline to 18 mo | 0.35 | 0.31 | 0.87 | 0.85 | 0.92 | 0.90 | 0.59 | 0.53 |
| (95% confidence interval) | (0.28, 0.43) | (0.23, 0.41) | (0.84, 0.90) | (0.82, 0.88) | (0.89, 0.94) | (0.86, 0.93) | (0.52, 0.65) | (0.46, 0.60) |
| 6 mo to 18 mo | 0.46 | 0.42 | 0.87 | 0.86 | 0.94 | 0.93 | 0.63 | 0.71 |
| (95% confidence interval) | (0.37, 0.55) | (0.33, 0.52) | (0.83, 0.90) | (0.82, 0.89) | (0.90, 0.95) | (0.89, 0.96) | (0.56, 0.69) | (0.66, 0.76) |
| 6 mo to 18 mo,[c] performers who provided appropriate effort only (n=215) | 0.42 | 0.46 | 0.86 | 0.85 | 0.93 | 0.91 | 0.62 | 0.71 |

[a] ICCs based on log-transformed data because of heteroscedasticity.

[b] ICCs for each pairwise comparison are provided in the first 6 rows for transparency reasons only; confidence intervals are omitted to improve clarity.

[c] Comparison to assess possible learning effects. The 95% confidence intervals were calculated using the bootstrap percentile method.[22]

value, the handgrip test takes very little time to administer and does not result in significant muscular fatigue. Our 95% limits of agreement (approximately –15 to 15) were much narrower than the previously reported –60 to 26 kg[24] (previous authors presented results as test 1 minus test 2, so a negative value reflected improvement). The larger variability in their study may be due to different methods, although the tests were conducted only 1 week apart and at exactly the same time of day. The previous authors tested participants 3 times at each session, allowed additional trials if force increased on the third trial, and then used the value from the trial with maximal force. We allowed only 2 trials and recorded the best value, disregarding any trials in which the performer pumped the handgrip. Even so, the limits of agreement from both studies reflect wide fluctuation in handgrip strength scores from test to test (±2 SDs represents a possible range of 30 kg of change for a variable, with a mean of 45 kg), making minor increases or decreases in strength difficult to identify.

Our results for the vertical jump test are also reassuring. Reported values for ICCs over 3 jumps on the same day were 0.98 using a stationary stance and 0.96 using the drop-step technique.[10] In addition, Burr et al[25] found both the Vertec and the Just Jump mat (Probotics, Inc, Huntsville, AL) highly reliable for both the squat jump and countermovement jump over 4 weeks (Vertec: ICC=0.98 to 0.99, respectively; Just Jump: ICC=0.99 for both). Other measures of lower body power have been reported as having higher ICCs, but they either require expensive force platform (or related) equipment (ICC>0.9)[26] or were measured over a period of several weeks (hop test for distance: ICC=0.86 to 0.96).[9,26] Although the hop test is even simpler than the Vertec in that it does not require extra equipment, it is a single-legged test with results that may be less transferable than jump height to other athletic movements.

Cirque du Soleil chose the 60-second jump test to measure anaerobic capacity. When we excluded the baseline data, the jump test showed relatively good reliability over the subsequent 3 time points (ICC=0.71). We believe this represents a learning effect of the administrators, because it is unlikely that the performers would have "learned" from a single test done 6 months previously, whereas the administrator would have applied the test to many performers during the first testing session. The gold standard for anaerobic capacity is the Wingate test (ICC=0.94 for average power, 0.83 for peak power),[27] but it is not easily conducted as a field test because it requires expensive equipment and skill to administer. Other common anaerobic fitness field tests with appropriate day-to-day reliabilities include the figure-8 hop test (ICC=0.92) and the up-and-down hop test (ICC=0.88). Although these tests take less time to administer, they have several disadvantages: both are single-legged tests, the figure-8 hop test takes more space (5 m×1 m versus 60 cm wide), and the up-and-down hop test requires a small box. Of note, CdS administrators stated that they were better able to fully focus on the performer's ability to complete the test properly if they used a countdown timer (in place of a stopwatch) to signal when time expired and a tally counter to keep track of correctly completed cycles.

Aerobic capacity was assessed using the modified Harvard step test. After we excluded the baseline data, the ICC was only marginally acceptable at 0.63. Advantages for the modified Harvard step test include the fact that it can be conducted anywhere because it requires minimal equipment and no electricity or calibration, is inexpensive, and only takes 8 minutes (5 minutes of exercise, 3 minutes of monitored recovery) to complete; in addition, the stepping skill takes little practice. Other options for measuring aerobic fitness include the 12-minute run test ($r=0.90$),[28] interval shuttle runs (ICCs=0.86 to 0.96 for men, 0.95 to 0.99 for women),[29] and 1-mile track walk test (also known as the Rockport Fitness Test, $r=0.93$).[28] Although these protocols have higher reliability, important disadvantages include the necessity of a large space or track, more time to conduct, and the need for constant internal motivation because the performer must be able to pace himself or herself in order to complete the entire assessment.

Of the 6 tests examined, only the dynamic balance assessment had poor reliability throughout the testing sessions. Although reliability increased somewhat over time, it never reached an acceptable limit of 0.60. Our results are consistent with those reported in the original description of the test by

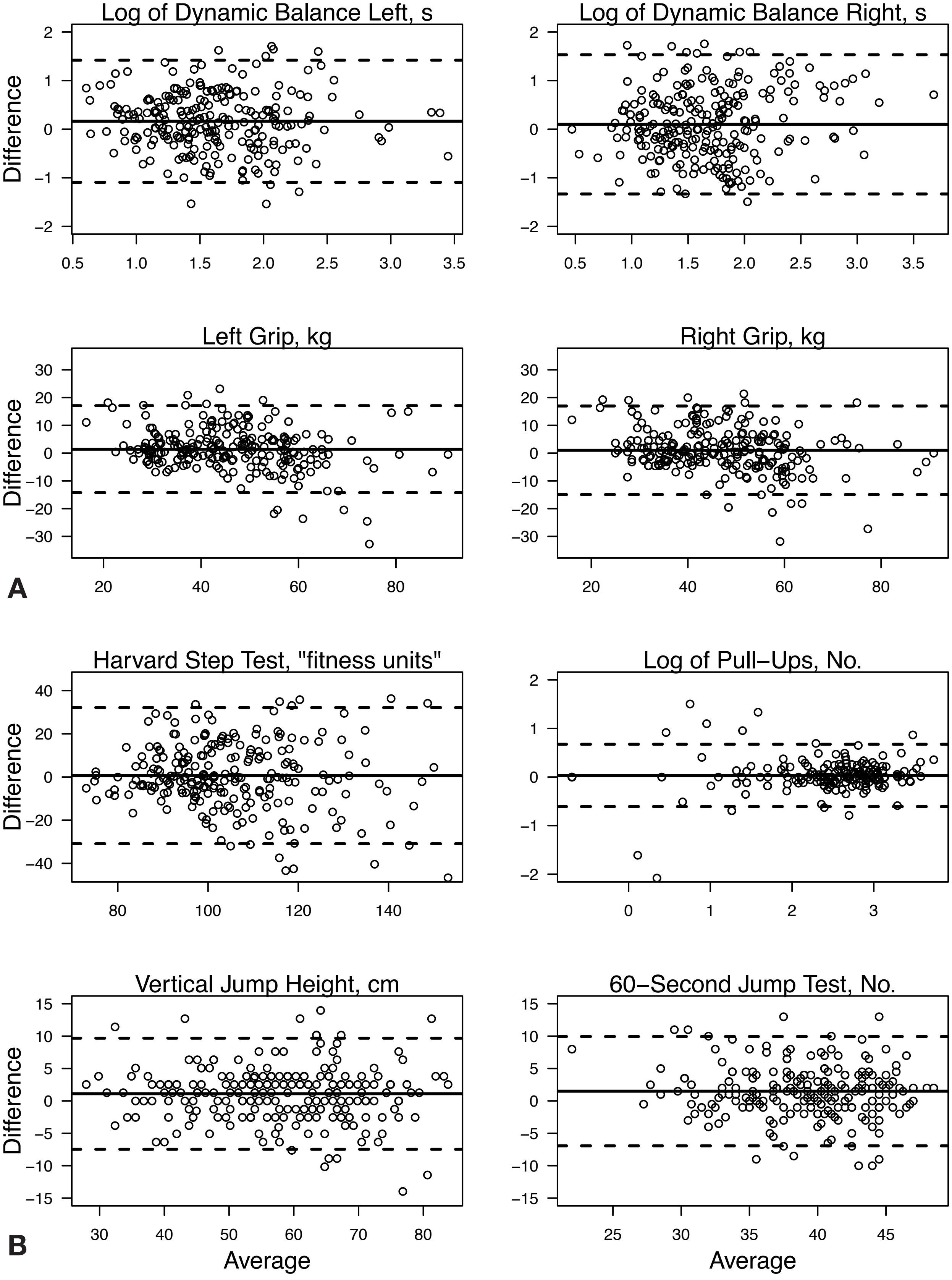

**Figure 2. Scatterplots represent the limits of agreement (difference between 2 tests plotted against the average of the 2 tests) for each of the physical capacity assessments conducted. The analysis is limited to the comparison of 18-month and 6-month data because our preliminary data indicated that baseline testing was not reliable due to administrator issues (see text). The solid lines represent the overall means for all performers at all 4 time points. The dotted lines above and below the center solid line represent the upper and lower limits of agreement.**

Emery et al[14] over a short period of time (ICC [3,1] = 0.46) and better than others[13] (ICC [3,1] = 0.00 to 0.02). Although our test had poor reliability, other current options also have important limitations. The single-legged balance test conducted on a firm surface rather than a foam pad (known as static balance) with eyes closed has greater reliability (ICC [3,1] = 0.69,[14] 0.25 to 0.58[13]) but the log-transformed 95% limits of agreement were wider (static = 0.28 to 3.2, dynamic = 0.48 to 2.29).[14] Also, the test takes longer to complete on average (geometric mean = 25.4 seconds [range = 3.8 to 148 seconds] for static balance and 5.3 seconds [range = 2.4 to 19.6 seconds] for dynamic balance).[14] The Star Excursion Balance Test (ICC = 0.78 to 0.96)[30] requires practice sessions before data are recorded (ie, it is more time consuming)[31] and is affected by leg length, height, foot type, and range of motion (factors that are less relevant if one is interested in intraindividual changes).[32]

Our study had a number of potential limitations. First, participants may not have provided maximal effort despite specific motivational strategies (eg, promoting a competitive spirit, direct verbal motivation during the test, and education about the importance of the test). However, this was probably not an important problem in our study because test administrators documented the level of effort (*appropriate* for 215/238 performers), and the ICC for the 215 performers who always showed appropriate effort was essentially the same as that of the entire group. Second, the test battery was conducted in the context of a performing arts company and, therefore, the timing varied from show to show, which could affect fatigue levels. For example, some performers' testing was during the early afternoon and concluded 2 hours before a performance, whereas other performers were usually tested after the show. In general, each show's performers were tested at the same time across all time points, but some exceptions existed. In addition, tests were generally but not always conducted on the last workday of the week. This lack of 100% consistency for the timing of test administration is a reality that must be understood when testing occurs outside a research environment. Thus, the lack of consistency was a strength of our study because it reflects real-world situations. That said, although we do not have evidence for or against our beliefs, we believe the magnitude of such a fatigue effect should theoretically reduce only reliability related to strength-endurance tests (ie, Harvard step test, pull-ups, 60-second jump test) and would not significantly affect balance, handgrip, or vertical jump tests. Third, turnover of administrative staff (as occurs in any real-world context) would be expected to reduce homogeneity of the methods, even though training videos and virtual meetings were available for new administrators. Fourth, we did not take into consideration previous medical history. Anyone who was performing at full duty was expected to participate in the testing sessions, but some volunteers may have had minor injuries that did not limit activity. That said, we specifically excluded performers who provided a submaximal effort if this was considered appropriate because of an ongoing health condition. Finally, we did not specifically ask whether performers were currently involved in any individual maintenance conditioning programs. However, no clinically relevant increases were noted in the mean scores for the majority of tests, which suggests that any training effects would be minimal.

In conclusion, our results suggest that from a group perspective, the modified Harvard step test, handgrip, vertical jump, pull-ups, and 60-second jump test were all reliable in the context of fitness testing as used in sport and occupational settings; the dynamic balance test was not. However, a learning effect was apparent for the Harvard step test and 60-second jump test. Also, although our ICC results indicated that this particular battery of tests may prove useful to identify differences among individuals, the limits-of-agreement results reflect important limitations in detecting changes in conditioning or deconditioning of an individual over time. Based on these results, CdS decided to continue to use these tests to develop more normative data and will follow up if trends appear over more than 2 periods or if the change in fitness tests correlates with changes in other aspects of performance-related measures. Clinicians should therefore realize that in addition to considerations of space, equipment, and time, it is necessary to plan how they intend to use the information generated when they select which fitness tests are most appropriate for their needs.

## REFERENCES

1. Williford HN, Duey WJ, Olson MS, Howard R, Wang N. Relationship between fire fighting suppression tasks and physical fitness. *Ergonomics*. 1999;42(9):1179–1186.
2. Rhodes EC, Farenholtz DW. Police Officer's Physical Abilities Test compared to measures of physical fitness. *Can J Sport Sci*. 1992;17(3): 228–233.
3. Magal M, Smith RT, Dyer JJ, Hoffman JR. Seasonal variation in physical performance-related variables in male NCAA Division III soccer players. *J Strength Cond Res*. 2009;23(9):2555–2559.
4. Health related physical testing and interpretation. In: Whaley MH, Brubaker PH, Otto RM, eds. *ACSM's Guidelines for Exercise Testing and Prescription*. 7th ed. Philadelphia, PA: Lippincott, Williams & Wilkins; 2006:55–92.
5. Tse MA, McManus AM, Masters RS. Development and validation of a core endurance intervention program: implications for performance in college-age rowers. *J Strength Cond Res*. 2005;19(3):547–552.
6. Clark MA, Russell AM. Integrated performance profile. In: Wittkop RT, ed. *Optimum Performance Training for the Performance Enhancement Specialist Course Manual*. Calabasas, CA: National Academy of Sports Medicine; 2001:115–186.
7. Hoffman JR, Tenenbaum G, Maresh CM, Kraemer WJ. Relationship between athletic performance tests and playing time in elite college basketball players. *J Strength Cond Res*. 1996;10(2):67–71.
8. Jinzhou Y, Fu Y, Zhang R, Li X, Shan G. The reliability and sensitivity of indices related to cardiovascular fitness evaluation. *Kinesiology*. 2008;40(2):139–146.
9. Ageberg E, Zatterstrom R, Moritz U. Stabilometry and one-leg hop test have high test-retest reliability. *Scand J Med Sci Sports*. 1998;8(4): 198–202.
10. Brodt V, Wagner DR, Heath EM. Countermovement vertical jump with drop step is higher than without in collegiate football players. *J Strength Cond Res*. 2008;22(4):1382–1385.
11. Peolsson A, Hedlund R, Oberg B. Intra- and inter-tester reliability and reference values for hand strength. *J Rehabil Med*. 2001;33(1):36–41.
12. Kamimura T, Ikuta Y. Evaluation of grip strength with a sustained maximal isometric contraction for 6 and 10 seconds. *J Rehabil Med*. 2001;33(5):225–229.
13. Schneiders AG, Sullivan SJ, Gray AR, Hammond-Tooke GD, McCrory PR. Normative values for three clinical measures of motor performance used in the neurological assessment of sports concussion. *J Sci Med Sport*. 2010;13(2):196–201.
14. Emery CA, Cassidy JD, Klassen TP, Rosychuk RJ, Rowe BB. Development of a clinical static and dynamic standing balance measurement tool appropriate for use in adolescents. *Phys Ther*. 2005;85(6):502–514.
15. Broglio SP, Ferrara MS, Macciocchi SN, Baumgartner TA, Elliott R. Test-retest reliability of computerized concussion assessment programs. *J Athl Train*. 2007;42(4):509–514.
16. Emery CA, Cassidy JD, Klassen TP, Rosychuk RJ, Rowe BH. Effectiveness of a home-based balance-training program in reducing sports-related

injuries among healthy adolescents: a cluster randomized controlled trial. *CMAJ*. 2005;172(6):749–754.

17. Katzmarzyk PT, Craig CL. Musculoskeletal fitness and risk of mortality. *Med Sci Sports Exerc*. 2002;34(5):740–744.
18. Rutherford WJ, Corbin CB. Validation of criterion-referenced standards for tests of arm and shoulder girdle strength and endurance. *Res Q Exerc Sport*. 1994;65(2):110–119.
19. Hoffman JR, Kang J. Evaluation of a new anaerobic power testing system. *J Strength Cond Res*. 2002;16(1):142–148.
20. Side-step test. Topend Sports Network Web site. http://www.topendsports.com/testing/tests/sidestep.htm. Accessed February 9, 2011.
21. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.
22. De Angelis D, Young GA. Bootstrap method. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. 2nd ed. Hoboken, NJ: Wiley; 2005. doi:10.1002/0470011815.b0470011812a0470014005. Accessed March 2009.
23. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–160.
24. Essendrop M, Schibye B, Hansen K. Reliability of isometric muscle strength tests for the trunk, hands and shoulders. *Int J Ind Ergon*. 2001;28(6):379–387.
25. Burr JF, Jamnik VK, Dogra S, Gledhill N. Evaluation of jump protocols to assess leg power and predict hockey playing potential. *J Strength Cond Res*. 2007;21(4):1139–1145.
26. Gustavsson A, Neeter C, Thomee P, et al. A test battery for evaluating hop performance in patients with an ACL injury and patients who have undergone ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc*. 2006;14(8):778–788.
27. Kerksick CM, Wilborn CD, Campbell BI, et al. Early-phase adaptations to a split-body, linear periodization resistance training program in college-aged and middle-aged men. *J Strength Cond Res*. 2009;23(3):962–971.
28. Noonan V, Dean E. Submaximal exercise testing: clinical application and interpretation. *Phys Ther*. 2000;80(8):782–807.
29. Lemmink KA, Visscher C, Lambert MI, Lamberts RP. The interval shuttle run test for intermittent sport players: evaluation of reliability. *J Strength Cond Res*. 2004;18(4):821–827.
30. Bressel E, Yonker JC, Kras J, Heath EM. Comparison of static and dynamic balance in female collegiate soccer, basketball, and gymnastics athletes. *J Athl Train*. 2007;42(1):42–46.
31. Kinzey SJ, Armstrong CW. The reliability of the Star-excursion test in assessing dynamic balance. *J Orthop Sports Phys Ther*. 1998;27(5):356–360.
32. Gribble PA, Hertel J. Considerations for normalizing measures of the Star Excursion Balance Test. *Meas Phys Educ Exerc Sci*. 2003;7(2):89–100.

*Address correspondence to Ian Shrier, MD, PhD, Centre for Clinical Epidemiology and Community Studies, Jewish General Hospital, 3755 Cote Ste-Catherine Road, Montreal, QC H3T 1E2, Canada. Address e-mail to ian.shrier@mcgill.ca.*