**Reliability of five trunk flexion and extension endurance field-based tests in high school-aged adolescents: ISQUIOS Programme**

**Running title:** Reliability of trunk endurance field-based tests

María Teresa Martínez-Romero [a,b], Francisco Ayala [b,c], Alba Aparicio-Sarmiento [a,b,*], Mark De Ste Croix [b,c], Pilar Sainz de Baranda [a,b]

[a] Department of Physical Activity and Sport, Faculty of Sport Sciences, Regional Campus of International Excellence "Campus Mare Nostrum", University of Murcia, San Javier 30720, Murcia, Spain

[b] Sports and Musculoskeletal System Research Group (RAQUIS), University of Murcia, Murcia 30100, Spain

[c] School of Sport and Exercise, Exercise and Sport Research Centre, University of Gloucestershire, Gloucester GL2 9HW, United Kingdom

*** Corresponding author:**

Email address: alba.aparicio@um.es (A., Aparicio-Sarmiento)

**Abstract**

This study aimed to explore the inter-session reliability of the measures obtained from 2 trunk extension (Biering-Sorensen and Dynamic Extensor Endurance (DEE) tests) and 3 trunk flexion (Ito, Side Bridge and Bench Trunk Curl-Up (BTC) tests) endurance field-based tests in adolescents by sex and age. A total of 208 (males, n = 100; females, n = 108) adolescents (ranging from 12 to 18 years) performed all the field-based tests on 2 separate testing sessions, 7-days apart. The inter-session reliability scores were explored for the total sample and by sex and age groups through relative reliability (intraclass correlation coefficient (ICC)), inter-session differences (systematic bias) and precision of measurements (i.e. absolute reliability)

(standard error of measurement expressed as a percentage of the mean score ($CV_{TE}$) and minimal detectable change ($MDC_{95}$)). The sensitivity of each test was also assessed through the smallest worthwhile percentage change (SWC). No relevant sex and age groups differences were found for either test-retest reliability or sensitivity in each test, so the grouped scores were considered as generalizable for this cohort of high school-aged adolescents. Most of the trunk endurance measures demonstrated acceptable relative reliability (ICCs ranged from 0.75 to 0.94). However, significant inter-session differences were identified for measures from the DEE and BTC tests. Likewise, the precision of the measurement of each field-based test was poor ($CV_{TE}$ ranged from 11.3 to 24.4%) with the $MDC_{95}$ revealing that changes higher than 42% for trunk extension endurance tests and 31.4% for trunk flexion endurance tests after an intervention are required to indicate a significant change above measurement error. All tests were sensitive enough to detect moderate to large changes in trunk muscle endurance. Therefore, the findings from this study indicate that only the BTC test demonstrates acceptable inter-session reliability (ICC > 0.9, $CV_{TE}$ ~ 10%, $MDC_{95}$ ~ 30%) to monitor the changes in trunk endurance scores that may be expected in adolescents after performing an intervention program. The use of supervised familiarization sessions before performing the tests and strong encouragement to perform a maximal effort in each test may be helpful strategies to improve the reliability scores.

**Keywords:** assessment; core endurance; precision of measurement; youth

## Introduction

It has been suggested that deficits in trunk extensor and flexor endurance and imbalances between trunk muscle groups may have short and long-term negative consequences in low back health[1,2] and athlete movement competency.[3,4] This circumstance has led to the field-based

assessment of trunk muscle endurance becoming common practice during childhood and adolescence, mainly in educational (physical education (PE) classes) and sport settings.

Some field-based tests have been described to assess trunk extensor and flexor muscle endurance, which may be grouped into two main categories:

a) Isometric trunk extension (e.g.: Biering-Sorensen (BS) test,[5] Prone Isometric Chest Raise (PICR) test[6] and Prone Double Straight-leg Raise (PDSR) test[7]) and flexion (e.g.: Ito test,[6] Flexor Endurance (FE) test,[8] Isometric Trunk Flexion endurance (ITF) test,[9] Plank Isometric Hold (PIH) test[10] and Side Bridge right (SB-R) and left (SB-L) test[8]) endurance tests, which involve maintaining a position against gravity for as long as possible.

b) Dynamic trunk extension (e.g.: Dynamic Extensor Endurance (DEE) test[11]) and flexion (e.g.: Bench Trunk Curl-Up (BTC) test,[12] Partial Curl-Up (PCU) test,[13] Curl-Up (CU) test[14]) endurance tests, which consist of performing as many repetitions as possible in a given time or with a certain cadence until exhaustion.

These field-based tests have been considered operationally valid by medical (ACSM), sport (Swiss Olympic Medical Centre) and educational (Cooper Institute) organizations to assess trunk muscles endurance based on anatomical knowledge and findings presented in electromyographic[15–17] and biomechanical[18–20] studies. It should be highlighted that it has not been described in the literature a single (i.e. all-out) field-based test able to quantify simultaneously the endurance of all trunk muscles (e.g., multifidus, transversus abdominis, external and internal abdominal obliques, erector spinae, quadratus lumborum, and rectus abdominis). Therefore, a comprehensive assessment of trunk endurance capability is required for each muscle group (e.g. flexors and extensors), by selecting at least one isometric and dynamic test.[18,21]

Reliability is a technical property of a measure or test that provides information regarding the consistency and reproducibility of given values in repeated trials.[22] The degree of reliability in a measure may be affected by the inter- and intra-individual variability in its scores within the sample object of study.[23] Consequently, reliability is a population dependent property (e.g.: children and adolescents, adults, athletes). Large inter-individual differences and fluctuations over short-time periods in strength and endurance scores have been documented in youth[24] and attributed, among other factors, to periods of rapid changes in growth and maturation and fluctuations in their mood state (e.g.: inter-day differences in the psychological readiness to perform a maximal effort to exhaustion).[22,25] Therefore, before promoting the use of these trunk endurance field-based tests in youth, the reliability of their measures must be confirmed in this population.[22] A recently published meta-analysis of reliability of the measures obtained from trunk extension endurance field-based tests concluded that, in terms of inter-session reliability, there is no compelling evidence that supports their use in children and adolescents.[26] In particular, only 3 studies were identified that provided inter-session reliability scores for the measures obtained from the BS and DEE in children[9,27] and adolescents[28] with all of them reporting intra-class correlation coefficient (ICC) scores higher than 0.80. Concerning the reliability of measures obtained from trunk flexion endurance field-based tests, the evidence available is also very limited in youth. Only 1 study has explored the inter-tester reliability of the dynamic endurance measure obtained from the CU test in children (10-12 years old),[29] whereas 5 studies have determined the inter-session reliability of the measures from PIH,[30] ITF,[9] isometric PCU[28] and BTC[25] tests in children and adolescents, showing ICC scores higher than 0.75.

Another limitation of the literature is that most of the studies that have explored the reliability of the trunk endurance field-based tests,[9,27–30] although not all,[25] have exclusively used the ICC as a criterion of reliability. The use of the ICC as the sole statistical outcome of reliability, apart

from being affected by sample heterogeneity, only provides information regarding how well the observed value retains the true rank order of subjects but does not allow the quantification of either the extent of the measurement error, the presence of systematic bias or the minimum change needed for a specific outcome to consider that an improvement or decrease after an intervention program may be real or true (out of the random error threshold). Therefore, contemporary statistical approaches in which the most powerful statistical methods were included, such as the typical (random) percentage error and the minimal detectable change at a 95% confidence interval ($MDC_{95}$), could be more useful in practical settings and for clinical goals.

Therefore, the purpose of the present study was to explore the inter-session reliability of the measures obtained from 2 trunk extension (BS and DEE tests) and 3 trunk flexion (Ito, SB [R and L] and BTC tests) endurance field-based tests using a contemporary statistical approach in high school-aged adolescents by sex and age. The null hypothesis is that unlike dynamic field-based tests, isometric trunk extensor and flexor muscle endurance measures would show acceptable (for clinical purposes and goal setting) and stronger reliability than dynamic measures, independent of the participants' age-based grade, due primarily to their easy assessment procedures.[25,31,32]

## Materials and methods

### *Participants*

A total of 241 adolescents were initially invited from 3 different high schools of the Region of Murcia (Spain) to participate in this study (convenience sample). The exclusion criteria were: a) known medical problems or episodes of low back pain over the last 3 months (reported by the PE teachers), b) not having provided the required signed written informed consent (by both the parents/guardians and students) before the start of the study, c) missing 1 testing session

during the data collection phase or d) involvement in structured strength exercise programs during the time of the study. Participants were asked not to perform strenuous exercises in the 24 h before each assessment session.

A comprehensive verbal description of the nature and purpose of the study and the experimental risks was given to the students and their parents/guardians and PE teachers. The study was conducted according to the Declaration of Frontera and the protocol was fully approved by the Review Committee for Research Involving Human Subjects at the University of Murcia (Spain) (ID: 1920/2018).

Finally, a sample of 208 (age 14.4 ± 1.2 years (mean ± SD), range 12-18 years, 52.4% girls) high school students (age groups (n): 12 to 13 years (70), 14 years (64), 15 to 18 years (74)) completed this study (Table 1). Thirty-three students were removed from the initial sample of 241 adolescents based on the exclusion criteria (10 students (3 boys and 7 girls) reported a history of low back pain, 14 students (6 boys and 8 girls) did not provide the required signed informed consent before the start of the study and 9 students (3 boys and 6 girls) did not attend one or both testing sessions).

*Study design and procedure*

A test-retest design was used to determine the inter-session reliability (both for all participants pooled in the same data set, as well as separated by sex and age-group) of the trunk endurance measures obtained from 5 field-based tests during PE classes. Two field-based tests were selected to assess isometric (BS test)[5] and dynamic (DEE test)[11] trunk extensor endurance (Fig. 1 A and B) whereas 3 field-based tests were selected to assess isometric (Ito and SB-R and SB-L tests)[6,8] and dynamic (BTC test)[12] trunk flexor endurance (Fig. 1 C, D and E). The tests included in this study were selected because they involve minimal equipment at low cost, and are feasible for administration in high-school settings[33,34] (13,55).

**Table 1.** Anthropometric characteristics of the participants (mean ± SD) (n=208).

| | **Total sample** | | **Age groups** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **≤ 13 years** | | **14 years** | | **≥ 15 years** | |
| | M | F | M | F | M | F | M | F |
| Sample | 100 | 108 | 34 | 36 | 39 | 25 | 27 | 47 |
| Age (years) | 14.4 ± 1.1 | 14.7 ± 1.4 | 13.2 ± 0.5 | 13 ± 0.5 | 14.5 ± 0.2 | 14.5 ± 0.2 | 15.8 ± 0.5 | 15.9 ± 0.6 |
| Body mass (kg) | 59.9 ± 14.7 | 56.2 ± 12 | 51.9 ± 10.1 | 55.6 ± 11.9 | 61.3 ± 14.1 | 56.5 ± 10.5 | 68.1 ± 15.8 | 60.6 ± 10.9 |
| Stature (cm) | 166.3 ± 9.5 | 159.4 ± 6.2 | 159.9 ± 8.7 | 156.7 ± 6.4 | 166.9 ± 7.1 | 161.9 ± 6.2 | 173.7 ± 7.8 | 160.3 ± 5.4 |
| BMI ($kg/m^2$) | 21.4 ± 4 | 22.8 ± 4.1 | 20.2 ± 3 | 22.6 ± 4.6 | 21.9 ± 4.3 | 21.6 ± 3.9 | 22.4 ± 4.4 | 23.5 ± 3.7 |

M = males, F = females, kg = kilogram, cm = centimetre, BMI = body mass index.

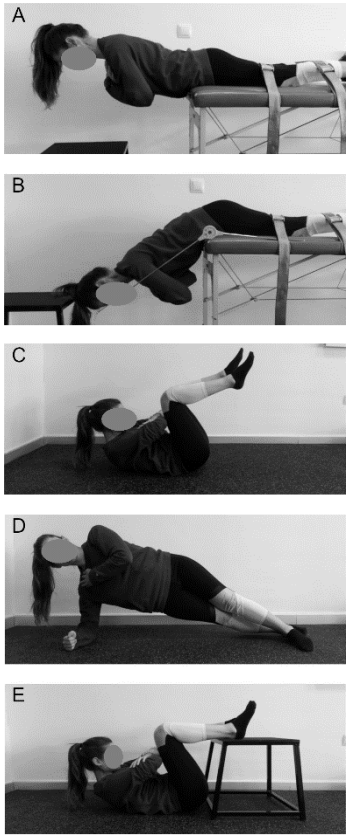


**Figure 1.** Trunk endurance field-based tests. A and B: trunk extension endurance field-based tests (BS test and DEE test). C, D, and E: trunk flexion endurance trunk field-based tests (Ito test, SB test, and BTC test). BS = Biering-Sorensen test, DEE = Dynamic Extensor Endurance test, SB = Side Bridge test, BTC = Bench Trunk Curl-Up test.

Since only 2 sessions of 60 min per age-based grade were provided by PE teachers from each high school, a time-efficient testing procedure was designed and 5 researchers were enrolled to enable the assessment of participants on 2 different occasions with a 7-days rest interval between them. The same protocol was consistently followed in all the testing sessions conducted in the 3 high schools that took part in this study. For each age grade, both testing sessions were administered at the same time of the day during PE classes and under the direct supervision of the same 5 researchers, who were sports science specialists with more than 5 years of experience in neuromuscular performance assessments. Each researcher was responsible during both testing sessions for the same field-based test. Furthermore, the participants' testing sequence and environmental factors were the same during the 2 testing sessions (Temperature: 22˚C, relative humidity: 50-60%).

At the start of the 2 testing sessions, all the participants received comprehensive instructions for the tests, and their questions regarding the protocols were answered. In each testing session, all participants completed first their usual warm-up, which was led by their PE teachers and consisted of 6-10 min of low-to moderate-intensity (self-perceived) running (including forward/backward movements and side-stepping) and general mobilization (i.e., arm circles, leg kicks) followed by 4-6 min of static stretching. Afterwards, the students were divided into 5 groups of 3-5 participants. Then, each researcher explained the procedure of the different tests to a group and instructed them to perform a minimum of 2 sets of 3 to 5 repetitions of each dynamic trunk endurance test and 3 sets of 5 seconds for the isometric endurance test. Once they had received the instructions, the researcher let them freely familiarize with the tests for 10 minutes. When the researchers verified that the participants had understood and freely practised the tests, participants randomly performed the 5 field-based trunk endurance tests with a 5-min rest between each test. Due to the above-mentioned time constraints, a circuit approach was used to carry out all the tests. Five different stations were set (one for each trunk endurance test). Each group of participants was randomly assigned to each station of the circuit in session 1 (for session 2, participants were assigned following the same order as in session 1). At each station, participants alternatively performed the test so that while one of them was carrying out the test the others were resting. After 8 min, groups were moved to their next station (clockwise) until all of them were completed.

An extendable goniometer (Lafayette Instrument Co, Lafayette, IN, USA) was used to ensure the correct joint position was maintained during the tests. Each tester had a digital stopwatch to quantify the time of the tests (CASIO HS-30W-N1V). During the performance of all field-based tests, participants were strongly encouraged verbally to maintain the position as long as possible or to perform the maximum number of repetitions as possible. Participants did not receive any feedback on performance until the end of the study.

*Trunk extensor endurance field-based tests*

*Biering-Sorensen test*

The isometric endurance of the trunk extensor musculature was assessed through the BS test.[5] The test started with the participant in a prone position with the lower body resting on a test bench and the anterior superior iliac spine aligned at the edge of the test bench. The lower body was attached to the test bench by 2 inextensible straps (knees and ankles). In the starting position, the upper body rested with both forearms placed on a chair. During the test, the upper body was maintained in a horizontal position (0 degrees of hip flexion) with arms crossed on the chest while holding the head in a neutral position (Fig. 1A). The test consisted of maintaining the trunk in the described position for as long as possible, until exhaustion, or until participants lost the correct position more than 3 times. A loss of the correct position during the execution of the test was identified when participants flexed their hips more than 10º (determined using extendable goniometer). The test duration was recorded in seconds.

*Dynamic Extensor Endurance test*

The dynamic endurance of the trunk extensor muscles was assessed through the DEE test.[11] Participants were located in the same position as the BS test. In the starting position, hip flexion of 45º was performed and both forearms rested on a chair. During the test, participants had to extend the trunk horizontally and then return to the initial position with arms crossed on the chest (Fig. 1B). Participants were asked to carry out the maximum repetitions possible in 60 seconds. Only those repetitions that were performed correctly were counted, that is, those in which the trunk was fully extending (horizontally), and in which the head touched the chair when flexing the hip. The hip flexion during the test was controlled through a static reference (extendable goniometer).

*Trunk flexor endurance field-based tests*

*Ito test*

The isometric endurance of the trunk flexor muscles was assessed through the Ito test.[6] Participants were placed in a supine position with hips and knees flexed 90° (extendable goniometer) and arms interlaced with hands grasping the opposite elbow. From this position, participants performed a trunk flexion ("curl-up") until they touched their thighs with their elbows, the scapulae did not touch the mat and the head was in a neutral position. The test consisted of maintaining this position for as long as possible, until exhaustion. The test ended when the scapulae came in contact with the mat, recording the test duration in seconds.

The original test was modified to normalize the range of motion to the participants' characteristics and thus avoid hip and lower back flexion ("sit-up").[19,20] For this, before starting the test, the participants were placed in the aforementioned initial position, and then the subject performed a trunk flexion until the scapulae did not touch the mat. From this position, the tester approached the participant's legs towards their elbows, until they came into contact (Fig. 1C). Then, the tester held the legs in this new position while the participant rested before starting the test. This leg position was maintained throughout the test.

*Side Bridge test*

The isometric endurance of the trunk lateral flexor musculature was assessed through the SB-R and SB-L test.[8] Participants were placed in a lateral position on their side (supported by either the dominant and non-dominant arm depending on the side tested) with legs extended. The participants were supported on their elbow and feet, the top foot was placed ahead of the lower foot (with 90º elbow flexion and the arm perpendicular to the mat) while bridging their hips off the mat to maintain an aligned body position. The uninvolved arm was held across the chest with the hand placed on the opposite shoulder (Fig. 1D). The test finished when the subject lost the aligned postural position, and the duration recorded in seconds. Both sides were tested with the dominant side always examined first.

*Bench Trunk Curl-Up test*

The dynamic endurance of the trunk flexor muscles was assessed through the BTC test.[12] Participants were placed in a supine position with hips and knees flexed at 90° (extendible goniometer) and resting on a bench. The arms were crossed with the hands grasping the opposite elbow (Fig. 1E). From this position, participants performed a trunk flexion ("curl-up") until they touched their thighs with their elbows, the scapulae did not touch the mat and the head was in a neutral position and then returned to the initial position. Just like the Ito test, a modification of the original test was performed to avoid hip and lower back flexion ("sit-up"), approaching the participant's legs towards their elbows, until they came into contact. The test consisted of performing the maximum number of repetitions possible in 2 minutes. Only those repetitions that were performed correctly were counted, that is, those in which the elbows touched the thighs in the flexing of the trunk, and in which the head touched the mat when lowering the trunk.

*Statistical Analyses*

Data are presented as mean ± SD. The distribution of each endurance measure was examined with the Shapiro-Wilk normality test and all measures were shown to be normally distributed.

In line with the current consensus regarding the determination of reliability in human performance-based studies, the following three aspects were assessed for each test using grouped and grade-specific measures:[33,34] 1) relative reliability, 2) presence (or not) of systematic bias between testing sessions and 3) precision of measurements (absolute reliability). Furthermore, the sensitivity of each test was also assessed.[35] The relative reliability was examined by the intra-class correlation coefficient ($ICC_{3,1}$) and an ICC >0.70 was considered acceptable.[36]

The assessment of systematic bias between testing sessions was carried out via the Bayesian paired t-test (with a Cauchy distribution with spread r set to 0.707). The $BF_{10}$ was interpreted using the evidence categories suggested by Lee and Wagenmakers:[37] <1/100 = extreme

evidence for $H_0$, from 1/100 to 1/30 = very strong evidence for $H_0$, from 1/30 to 1/10 = strong evidence for $H_0$, from 1/10 to 1/3 = moderate evidence for $H_0$, from 1/3 to 1 anecdotal evidence for $H_0$, from 1 to 3 = anecdotal evidence for $H_1$, from 3 to 10 = moderate evidence for $H_1$, from 10 to 30 = strong evidence for $H_1$, from 30 to 100 = very strong evidence for $H_1$, >100 extreme evidence for $H_1$. The median and the 95% central credible interval (CI) of the posterior distribution of the standardized effect size (δ) (i.e. the population version of Cohen's d) were also calculated for each of the paired-comparisons carried out. Magnitudes of the posterior distribution of the standardized effect size were classified as: trivial (<0.2), small (0.2 – 0.6), moderate (0.6 – 1.2), large (1.2 – 2.0) and very large (2.0 – 4.0).[38] Only those pairwise comparisons that showed at least strong evidence for supporting the alternative hypothesis ($BF_{10}$ >10), an error percentage <10 (which indicates great stability of the numerical algorithm that was used to obtain the result) and δ >0.6 (at least moderate) were considered robust to describe significant differences.

A Bland-Altman plot was built for each trunk endurance measure to graphically show mean bias and 95% limits of agreement. Heteroscedasticity was assessed using a Bayesian correlation coefficient (Pearson´s rho) between the means of the participant's test and retest scores and the absolute differences between the participant's test and retest scores.[22] To qualitatively interpret the size of the Bayesian correlation coefficients, the thresholds defined by Hinkle, Wiersma & Jurs[39] for Behavioural Sciences were followed: from 0 to 0.3 = negligible correlation, from 0.3 to 0.5 = low correlation, from 0.5 to 0.7 = moderate, from 0.7 to 0.9 = high, from 0.9 to 1 = very high. Only correlations higher than 0.5 (at least moderate) were considered relevant for these sub-analyses.

The precision of measurement was determined using the typical percentage error and the minimal detectable change at a 95% confidence interval ($MDC_{95}$) using the Hopkins' spreadsheet.[40] The typical percentage error (coefficient of variation ($CV_{TE}$)) was calculated

using the log-transformed data via the following formula: 100 ($e^{s} - 1$), where s is the typical error of measurement (TEM) (SD of the difference between testing session 1 and testing session 2 divided by $\sqrt{2}$). Logarithmic transformations of the data were performed and used to reduce the possible heteroscedasticity of the raw data.[41] To interpret the $CV_{TE}$ values, the current study used the arbitrary value suggested by Weir and Vincent[42] and Hopkins[22] with an analytical goal of 10% or below to consider a test as demonstrating good inter-session reliability. The $MDC_{95}$ was calculated as follow: $CV_{TE} \times 1.96 \times \sqrt{2}$.

The sensitivity of each test was assessed while comparing the smallest worthwhile percentage change (SWC) with the $CV_{TE}$. The SWC was determined by multiplying the pure between-testing sessions SD by 0.2 ($SWC_{0.2}$), which corresponds to a small effect, 0.6 ($SWC_{0.6}$), which corresponds to a moderate effect and 1.2 ($SWC_{1.2}$), which corresponds to a large effect. If the $CV_{TE}$ was lower than the SWC, the test was rated as "good"; if the $CV_{TE}$ was similar to the SWC, the rating was "OK"; and if the $CV_{TE}$ was higher than the SWC, a rating of "marginal" was given.[35]

Statistical analysis was performed using the JASP computer software Version 0.11.1 (JASP Team, Amsterdam, The Netherlands) and the online Hopkins' spreadsheet (www.sportsci.org).

**Results**

The descriptive statistics and reliability values for the trunk flexion and extension endurance measures obtained from the 5 field-based tests selected are shown in tables 2 and 3, respectively, for the total sample and by sex and age groups.

The relative reliability scores (i.e. $ICC_{3,1}$) found in this study for all the trunk endurance measures were higher than 0.7 (except the ICC scores for the DEE test in the group of boys,

**Table 2.** Descriptive statistics (mean ± SD) and reliability scores (mean and 90% confidence interval) of the trunk extension endurance measures obtained from the 2 field-based tests selected.

| | Total sample | | | Age group | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ≤ 13 years | | 14 years | | ≥ 15 years | |
| Sex | Total (n = 208) | M (n = 100) | F (n = 108) | M (n = 34) | F (n = 36) | M (n = 39) | F (n = 25) | M (n = 27) | F (n = 47) |
| BS test (s) | | | | | | | | | |
| - Testing session 1 | 132.5 ± 54.6 | 144.1 ± 61.2 | 121.8 ± 45.4 | 150.4 ± 71.1 | 135.1 ± 48.5 | 137.8 ± 49.4 | 117.8 ± 45.2 | 147.7 ± 69.5 | 115.5 ± 42.9 |
| - Testing session 2 | 130.5 ± 52.0 | 137.4 ± 57.1 | 124.1 ± 46.4 | 154.6 ± 67.6 | 131.9 ± 52.5 | 129.8 ± 47.1 | 126.8 ± 48.1 | 132.2 ± 59.5 | 117.7 ± 41.6 |
| - Systematic bias (%) | -1.3 (-4.0 to 1.6) | -4.6 (-8.6 to -0.4) | 1.9 (-1.9 to 5.8) | 3.1 (-6.4 to 13.5) | -4.4 (-11.6 to 3.5) | -6.3 (-11.2 to -1) | 8.7 (0.9 to 17.2) | -9.2 (-16.6 to -1) | 2.5 (-2.8 to 8.1) |
| - $ICC_{3,1}$ | 0.86 (0.82 to 0.89) | 0.86 (0.80 to 0.9) | 0.86 (0.8 to 0.9) | 0.86 (0.71 to 0.93) | 0.86 (0.73 to 0.93) | 0.87 (0.77 to 0.93) | 0.90 (0.8 to 0.96) | 0.89 (0.78 to 0.95) | 0.86 (0.76 to 0.91) |
| - $CV_{TE}$ (%) | 16 (14.6 to 17.9) | 16.5 (14.4 to 19.4) | 15.6 (13.4 to 17.9) | 19.3 (15 to 27.3) | 16.9 (13.4 to 23.2) | 13.7 (11.3 to 17.7) | 14.2 (11.1 to 20.2) | 17 (13.3 to 24) | 14.2 (11.8 to 18.1) |
| - $MDC_{95}$ (%) | 44.5 (40.4 to 49.6) | 45.7 (39.9 to 53.8) | 43.2 (37.1 to 49.6) | 53.4 (41.6 to 75.6) | 46.8 (37.1 to 64.2) | 38 (31.2 to 49) | 39.4 (30.6 to 55.9) | 47.1 (36.7 to 66.46) | 39.5 (32.7 to 50.1) |
| DEE test (rep) | | | | | | | | | |
| - Testing session 1 | 39.0 ± 10.6 | 43.9 ± 10.5 | 34.6 ± 8.6 | 41.6 ± 8.7 | 35.5 ± 8.8 | 46 ± 11.4 | 36.7 ± 10.4 | 43.4 ± 10.6 | 33.1 ± 8.8 |
| - Testing session 2 | 43.3 ± 12.6 | 49.5 ± 12.6 | 37.5 ± 9.5 | 46.2 ± 13.9 | 37.1 ± 10.1 | 52.2 ± 12.1 | 39.5 ± 8.2 | 49.5 ± 11.7 | 37.1 ± 9.7 |
| - Systematic bias (%) | 10 (7.1 to 12.9)* | 11.8 (7.5 to 16.3)* | 8.3 (4.6 to 12.3) | 8 (-1.1 to 18) | 4 (-1.6 to 9.9) | 13.4 (7.1 to 20.1)* | 7.8 (-0.1 to 16.4) | 13.7 (6.3 to 21.7)* | 12.1 (5.7 to 18.8) |
| - $ICC_{3,1}$ | 0.77 (0.72 to 0.82) | 0.69 (0.58 to 0.78) | 0.75 (0.67 to 0.82) | 0.61 (0.34 to 0.78) | 0.77 (0.62 to 0.87) | 0.74 (0.57 to 0.85) | 0.67 (0.36 to 0.85) | 0.74 (0.53 to 0.87) | 0.77 (0.64 to 0.86) |
| - $CV_{TE}$ (%) | 15.3 (14 to 17) | 15.8 (13.8 to 18.5) | 14.9 (13.1 to 17.3) | 19.5 (15.5 to 26.7) | 13.4 (11 to 17.5) | 14.2 (11.6 to 18.4) | 13.1 (10 to 19.4) | 13.9 (11 to 19.1) | 16.5 (13.7 to 20.8) |
| - $MDC_{95}$ (%) | 42.4 (38.7 to 47.2) | 43.8 (38.2 to 51.3) | 41.3 (36.3 to 47.9) | 54.1 (43 to 73.9) | 37.2 (30.3 to 48.5) | 39.3 (32.3 to 51) | 36.4 (27.8 to 53.8) | 38.5 (30.5 to 53.1) | 45.7 (38.1 to 57.5) |

*: there was at least a strong evidence ($BF_{10}$ >10) to support the alternative hypothesis ($H_1$: the presence of relevant inter-session differences) with at least a moderate effect size (δ >0.6). ICC = intraclass correlation coefficient, $CV_{TE}$ = typical percentage error, MDC = minimum detectable change, BS = Biering-Sorensen test, DEE = Dynamic Extensor Endurance test

**Table 3.** Descriptive statistics (mean ± SD) and reliability scores (mean and 90% confidence interval) of the trunk flexion endurance measures obtained from the 3 field-based tests selected.

| | Total sample | | | Age group | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ≤ 13 years | | 14 years | | ≥ 15 years | |
| Sex | Total (n = 208) | M (n = 100) | F (n = 108) | M (n = 34) | F (n = 36) | M (n = 39) | F (n = 25) | M (n = 27) | F (n = 47) |
| Ito test | | | | | | | | | |
| - Testing session 1 | 160.1 ± 142.5 | 164.5 ± 151.1 | 156.2 ± 135.5 | 128.5 ± 81.7 | 141.3 ± 122.1 | 188.3 ± 172.3 | 182.5 ± 187.1 | 166.2 ± 171.4 | 156 ± 128.1 |
| - Testing session 2 | 165.4 ± 147.4 | 170.7 ± 155.6 | 160.8 ± 140.8 | 138.8 ± 94.2 | 142.3 ± 134.5 | 194.9 ± 174.4 | 195.8 ± 186.2 | 167.8 ± 176.9 | 160.1 ± 132.2 |
| - Systematic bias (%) | 1.3 (-3.1 to 5.9) | 2.8 (-3.1 to 9.1) | 0 (-6.3 to 6.8) | 4 (-7.1 to 16.5) | -7.1 (-17.5 to 4.6) | 7.9 (-0.2 to 16.6) | 14.4 (-5.4 to 38.5) | -5.2 (-17.5 to 8.8) | -0.2 (-8.6 to 9) |
| - $ICC_{3,1}$ | 0.94 (0.92 to 0.95) | 0.95 (0.93 to 0.97) | 0.92 (0.89 to 0.95) | 0.92 (0.83 to 0.96) | 0.94 (0.88 to 0.97) | 0.98 (0.96 to 0.99) | 0.94 (0.84 to 0.98) | 0.94 (0.87 to 0.97) | 0.92 (0.87 to 0.95) |
| - $CV_{TE}$ (%) | 24.4 (21.9 to 27.5) | 21.8 (18.8 to 26.2) | 26.5 (23 to 31.4) | 21.6 (16.6 to 31.5) | 23.4 (18 to 33.8) | 17.9 (14.3 to 24) | 28 (20 to 48.1) | 26.9 (20.5 to 39.5) | 27.2 (22.7 to 34.3) |
| - $MDC_{95}$ (%) | 67.5 (60.8 to 76.1) | 60.4 (52.1 to 72.6) | 73.4 (63.7 to 87.1) | 59.8 (45.9 to 87.2) | 64.8 (50 to 93.6) | 49.5 (39.7 to 66.5) | 77.5 (55.4 to 133.4) | 74.5 (56.9 to 109.6) | 75.4 (62.8 to 95) |
| SB-R test | | | | | | | | | |
| - Testing session 1 | 57.3 ± 29.4 | 67.3 ± 28.9 | 48.7 ± 25.9 | 62.4 ± 30.1 | 53.3 ± 29.7 | 59.1 ± 23.9 | 41.9 ± 22.2 | 82.7 ± 29.2 | 48.6 ± 24.6 |
| - Testing session 2 | 56.4 ± 27.4 | 64.6 ± 28.1 | 49.6 ± 23.6 | 58.3 ± 31.6 | 54.1 ± 26.1 | 57.5 ± 20.2 | 43 ± 22.2 | 79.9 ± 28.9 | 49.4 ± 22.3 |
| - Systematic bias (%) | 1.7 (-1.7 to 5.3) | -3.8 (-6.9 to -0.6) | 7.5 (1.7 to 13.7) | -8.4 (-15.1 to -1.2) | 11.6 (-2.1 to 27.1) | 0 (-4.9 to 5.2) | 5.2 (-4.8 to 16.1) | -4.4 (-9 to 0.4) | 5.8 (-1.4 to 13.5) |
| - $ICC_{3,1}$ | 0.90 (0.87 to 0.92) | 0.94 (0.91 to 0.96) | 0.89 (0.81 to 0.9) | 0.94 (0.88 to 0.97) | 0.83 (0.7 to 0.91) | 0.93 (0.88 to 0.96) | 0.93 (0.84 to 0.97) | 0.94 (0.88 to 0.97) | 0.88 (0.8 to 0.93) |
| - $CV_{TE}$ (%) | 19.7 (17.9 to 22) | 12.8 (11.2 to 15) | 24.5 (21.5 to 28.6) | 15.3 (12 to 21.3) | 33.2 (26.5 to 44.8) | 12.4 (10.2 to 16.1) | 18.7 (14.4 to 27.1) | 10.3 (8.2 to 14) | 20.5 (17 to 25.8) |
| - $MDC_{95}$ (%) | 54.6 (49.5 to 60.9) | 35.5 (31 to 41.6) | 68 (59.7 to 79.2) | 42.3 (33.2 to 58.9) | 91.9 (73.4 to 124.2) | 34.4 (28.2 to 44.6) | 51.7 (39.8 to 75.1) | 28.5 (22.7 to 38.7) | 56.7 (47.2 to 71.4) |
| SB-L test | | | | | | | | | |
| - Testing session 1 | 57.3 ± 29.7 | 67.1 ± 27.8 | 50.1 ± 29.5 | 60.8 ± 29.4 | 58.9 ± 35.5 | 62.6 ± 28.3 | 40.8 ± 21.3 | 78.8 ± 22.6 | 48.1 ± 26.9 |
| - Testing session 2 | 55.5 ± 27.4 | 65.4 ± 24.8 | 48.4 ± 27.2 | 60.2 ± 25.6 | 56.9 ± 34.5 | 61.8 ± 24.2 | 39.1 ± 18.9 | 75.1 ± 23.2 | 46.7 ± 23.2 |
| - Systematic bias (%) | -1.2 (-4.3 to 2.1) | -0.7 (-4.4 to 3.1) | -1.2 (-6.3 to 4) | -0.1 (-8.4 to 8.8) | -2.9 (-11.4 to 6.3) | 3.1 (-3 to 9.7) | -2.4 (-13.9 to 10.6) | -5.6 (-10.2 to -0.7) | 0.5 (-7.1 to 8.7) |
| - $ICC_{3,1}$ | 0.92 | 0.93 | 0.90 | 0.90 | 0.93 | 0.94 | 0.87 | 0.92 | 0.88 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (0.90 to 0.94) | (0.89 to 0.95) | (0.85 to 0.93) | (0.81 to 0.95) | (0.86 to 0.96) | (0.89 to 0.97) | (0.72 to 0.94) | (0.85 to 0.96) | (0.81 to 0.93) |
| - $CV_{TE}$ (%) | 18.5 | 14.9 | 22.8 | 17.6 | 22.1 | 15.3 | 24.1 | 10.5 | 23.2 |
| | (16.8 to 20.7) | (13 to 17.4) | (20 to 26.5) | (13.8 to 24.5) | (17.8 to 29.5) | (12.5 to 19.9) | (18.5 to 35.3) | (8.3 to 14.2) | (19.3 to 29.3) |
| - $MDC_{95}$ (%) | 51.4 | 41.3 | 63.1 | 48.6 | 61.3 | 42.5 | 66.7 | 29 | 64.2 |
| | (46.6 to 57.4) | (36 to 48.2) | (55.4 to 73.4) | (38.1 to 68) | (49.3 to 81.7) | (34.7 to 55.2) | (51.2 to 97.8) | (23.1 to 39.4) | (53.4 to 81.1) |
| BTC test | | | | | | | | | |
| - Testing session 1 | 57.5 ± 18.8 | 58 ± 18.6 | 57 ± 19.1 | 57.9 ± 22.7 | 56.2 ± 20.3 | 52.6 ± 16.1 | 60.4 ± 21.4 | 65.1 ± 14.8 | 56.5 ± 17.7 |
| - Testing session 2 | 64.9 ± 19.8 | 65.1 ± 20 | 64.6 ± 19.6 | 62.2 ± 23.6 | 61.7 ± 21.4 | 62.5 ± 19.5 | 73.5 ± 21 | 71.7 ± 15.2 | 63.9 ± 17.2 |
| - Systematic bias | 13.6 | 12.8 | 14.4 | 7.8 | 10.4 | 19.3 | 23.9 | 10.3 | 14.5 |
| (%) | (11.3 to 15.9)* | (9.8 to 15.8)* | (10.9 to 18.1)* | (2.3 to 13.6) | (4.8 to 16.3)* | (14.5 to 24.3)* | (16.4 to 31.8)* | (5.7 to 15.1)* | (9.1 to 20.3)* |
| - $ICC_{3,1}$ | 0.90 | 0.92 | 0.88 | 0.94 | 0.9 | 0.94 | 0.94 | 0.89 | 0.86 |
| | (0.87 to 0.92) | (0.89 to 0.95) | (0.83 to 0.92) | (0.88 to 0.97) | (0.82 to 0.95) | (0.88 to 0.97) | (0.84 to 0.98) | (0.78 to 0.94) | (0.77 to 0.92) |
| - $CV_{TE}$ (%) | 11.3 | 10.2 | 12.4 | 11.2 | 12.1 | 9.5 | 8.8 | 8.6 | 13.1 |
| | (10.3 to 12.6) | (9 to 12) | (10.8 to 14.4) | (9 to 15.1) | (9.8 to 15.9) | (7.7 to 12.4) | (6.5 to 14) | (6.8 to 11.7) | (10.9 to 16.6) |
| - $MDC_{95}$ (%) | 31.4 | 28.3 | 34.2 | 31 | 33.4 | 26.2 | 24.4 | 23.7 | 36.9 |
| | (28.5 to 34.9) | (24.9 to 33.3) | (30 to 39.9) | (24.8 to 41.8) | (27.1 to 43.9) | (21.3 to 34.3) | (18.1 to 38.8) | (18.9 to 32.4) | (30.2 to 46.1) |

*: there was at least a strong evidence ($BF_{10}$ >10) to support the alternative hypothesis ($H_1$: the presence of relevant inter-session differences) with at least a moderate effect size ($\delta$ >0.6). ICC = intraclass correlation coefficient, $CV_{TE}$ = typical percentage error, MDC = minimum detectable change, SB-L = Side Bridge Left test, SB-R = Side Bridge Right test, BTC = Bench Trunk Curl-Up test.

in 13-year-old boys, and in 14-year-old girls) and hence, they may be considered as acceptable according to the thresholds previously reported in the literature.[36]

The separate Bayesian paired t-test analyses carried out to explore potential inter-session differences (systematic bias) only revealed the existence of at least a strong evidence in favour of the alternative hypothesis ($H_1$) with at least a moderate effect size ($\delta > 0.6$) in the trunk extension and flexion endurance measures obtained from their respective dynamic field-based tests (DEE and BTC tests). Bland-Altman plots illustrate the differences between the two testing sessions (y-axis) and the mean value of each of the paired measurements (x-axis) for each trunk endurance field-based test (Fig. 2 (grouped data) and supplementary files 1-4 (by sex and age groups)). Dashed lines illustrate the systematic bias and random error forming the 95% limits of agreement.

The heteroscedasticity coefficients for the measures obtained through the 5 trunk endurance field-based tests were not relevant for the grouped and by sex and age groups ($r < 0.5$) (except the heterogeneity correlation scores found for the DEE test ($r = -0.6$) in 13-year-old boys) (Fig. 2).

The $CV_{TE}$ scores obtained from the field-based tests for the total sample ranged from 15.3% (DEE test) to 18% (BS test) and from 11.3% (BTC test) to 24.4% (Ito test) for the trunk extensor and flexor muscles, respectively. The males' group had lower $CV_{TE}$ scores in comparison with the females' group in 4 of the 6 field-based tests administered, concretely in the trunk flexor endurance field-based tests. When analysing the $CV_{TE}$ scores by age groups and sex, the youngest boys ($\leq$ 13 years) generally presented higher $CV_{TE}$ percentages, unlike the girls, who presented higher percentages in the group over 15 years (Tables 2 and 3).

For the $MDC_{95}$, the total sample's scores ranged from 42.4% (DEE test) to 44.5% (BS test) and from 31.4% (BTC test) to 67.5% (Ito test) for the trunk flexion and extension endurance measures, respectively.

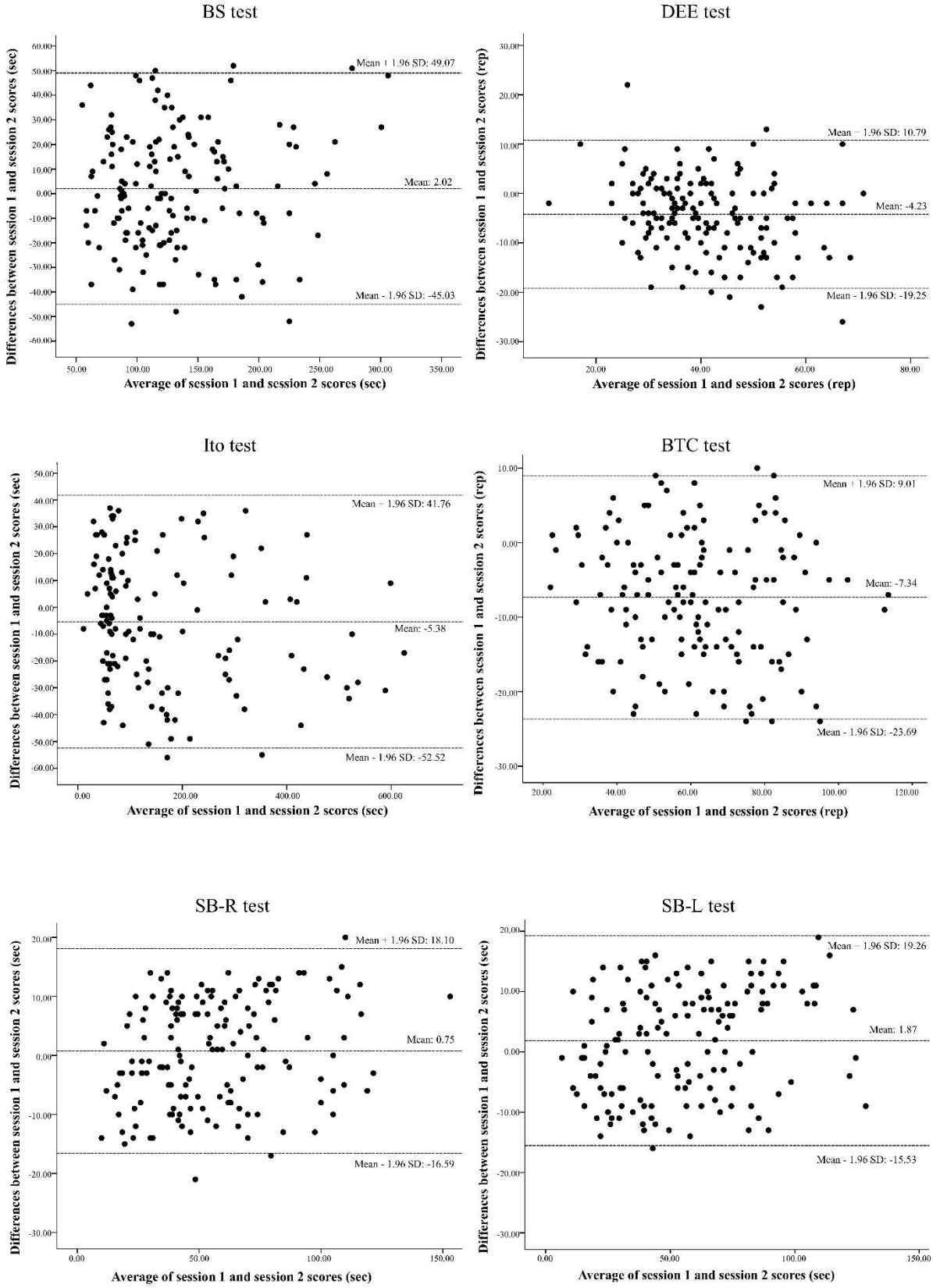


**Figure 2.** Bland-Altman plots for the trunk endurance field-based tests (grouped data (n = 208). Trunk extension endurance field-based tests (BS test and DEE test). Trunk flexion endurance trunk field-based tests (Ito test, SB

test, and BTC test). BS = Biering-Sorensen test, DEE = Dynamic Extensor Endurance test, SB = Side Bridge test, BTC = Bench Trunk Curl-Up test.

As displayed in Table 4, all field-based tests were sensitive to detect moderate and large changes (independent of the participants' sex and age group), but none of the tests was considered sensitive enough to detect small changes.

**Discussion**

The main purpose of this study was to examine the inter-session reliability of the trunk flexor and extensor endurance measures obtained from 5 common field-based tests during PE classes in a large sample of high school-aged adolescents by sex and age. The analyses of the inter-session reliability and sensitivity carried out with the grouped and sex and age group-specific trunk flexor and extensor endurance measures consistently reported similar results across both sexes and the 3 age groups. Therefore, both the relative and absolute reliability and also the sensitivity scores obtained from each field-based test using the measures from the whole sample of participants (i.e. grouped data set) may be considered as robust (based on the large sample size (n = 208)) and generalizable criteria to be used when assessing and monitoring trunk endurance in high school-aged adolescents (i.e. independent of the age group). Thus, the findings of the present study indicate that all trunk flexor and extensor endurance measures demonstrate acceptable (ICC > 0.7) relative reliability. However, significant inter-session differences (i.e. systematic bias) were found in the measures obtained from the DEE and BTC tests. Likewise, the precision of the measurement of each field-based test was poor ($CV_{TE}$ <10%) with the $MDC_{95}$ revealing that changes higher than 42% for trunk extension endurance tests and 31.4% for trunk flexion endurance tests after an intervention are required to indicate a significant change above measurement error. Finally, the sensitivity

**Table 4.** The sensitivity of the field-based trunk endurance tests to detect small, moderate and large changes. Rating of sensitivity is also provided

| | Total sample | | | Age group | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ≤ 13 years | | 14 years | | ≥ 15 years | |
| Sex | Total (n = 208) | M (n = 100) | F (n = 108) | M (n = 34) | F (n = 36) | M (n = 39) | F (n = 25) | M (n = 27) | F (n = 47) |
| **Trunk extension endurance field-based tests** | | | | | | | | | |
| BS test | | | | | | | | | |
| - $SWC_{0.2}$ | 7.5 (marginal) | 7.7 (marginal) | 7.2 (marginal) | 8.5 (marginal) | 7.6 (marginal) | 6.6 (marginal) | 7.9 (marginal) | 9 (marginal) | 6.5 (marginal) |
| - $SWC_{0.6}$ | 15 (OK) | 23.1 (good) | 21.6 (good) | 25.5 (good) | 22.8 (good) | 19.8 (good) | 23.7 (good) | 27 (good) | 19.5 (good) |
| - $SWC_{1.2}$ | 45 (good) | 46.2 (good) | 43.2 (good) | 51 (good) | 45.6 (good) | 39.6 (good) | 47.4 (good) | 54 (good) | 39 (good) |
| DEE test | | | | | | | | | |
| - $SWC_{0.2}$ | 5.4 (marginal) | 4.4 (marginal) | 4.9 (marginal) | 4.4 (marginal) | 4.6 (marginal) | 4.4 (marginal) | 3.4 (marginal) | 4.3 (marginal) | 5.6 (marginal) |
| - $SWC_{0.6}$ | 10.8 (good) | 13.2 (good) | 14.7 (OK) | 13.2 (marginal) | 13.8 (OK) | 13.2 (marginal) | 10.2 (marginal) | 12.9 (marginal) | 16.8 (OK) |
| - $SWC_{1.2}$ | 32.5 (good) | 26.4 (good) | 29.4 (good) | 26.4 (good) | 27.6 (good) | 26.4 (good) | 20.4 (good) | 25.8 (good) | 33.6 (good) |
| **Trunk flexion endurance field-based tests** | | | | | | | | | |
| Ito test | | | | | | | | | |
| - $SWC_{0.2}$ | 18.1 (marginal) | 19.1 (marginal) | 17.3 (marginal) | 13.2 (marginal) | 17.5 (marginal) | 22.6 (marginal) | 19.4 (marginal) | 19.7 (marginal) | 17 (marginal) |
| - $SWC_{0.6}$ | 36.2 (good) | 57.3 (good) | 51.9 (good) | 39.6 (good) | 52.5 (good) | 67.8 (good) | 58.2 (good) | 59.1 (good) | 51 (good) |
| - $SWC_{1.2}$ | 108.6 (good) | 114.6 (good) | 103.8 (good) | 79.2 (good) | 105 (good) | 135.6 (good) | 116.4 (good) | 118.2 (good) | 102 (good) |
| SB-R test | | | | | | | | | |
| - $SWC_{0.2}$ | 11.5 (marginal) | 9.8 (marginal) | 11.5 (marginal) | 11.2 (marginal) | 12.9 (marginal) | 8.9 (marginal) | 12 (marginal) | 7.6 (marginal) | 10.4 (marginal) |
| - $SWC_{0.6}$ | 23 (good) | 29.4 (good) | 34.5 (good) | 33.6 (good) | 38.7 (good) | 26.7 (good) | 36 (good) | 22.8 (good) | 31.2 (good) |
| - $SWC_{1.2}$ | 69 (good) | 58.8 (good) | 69 (good) | 67.2 (good) | 77.4 (good) | 53.4 (good) | 72 (good) | 45.6 (good) | 62.4 (good) |
| SB-L test | | | | | | | | | |
| - $SWC_{0.2}$ | 12.3 (marginal) | 10.1 (marginal) | 12.6 (marginal) | 9.9 (marginal) | 14 (marginal) | 11.7 (marginal) | 10.9 (marginal) | 6.7 (marginal) | 11.8 (marginal) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| - $SWC_{0.6}$ | 24.6 (good) | 30.3 (good) | 37.8 (good) | 29.7 (good) | 42 (good) | 35.1 (good) | 32.7 (good) | 20.1 (good) | 35.4 (good) |
| - $SWC_{1.2}$ | 73.8 (good) | 60.6 (good) | 75.6 (good) | 59.4 (good) | 84 (good) | 70.2 (good) | 65.4 (good) | 40.2 (good) | 70.8 (good) |
| BTC test | | | | | | | | | |
| - $SWC_{0.2}$ | 6.7 (marginal) | 7 (marginal) | 6.4 (marginal) | 8.4 (marginal) | 6.8 (marginal) | 6.9 (marginal) | 6.2 (marginal) | 4.5 (marginal) | 6.2 (marginal) |
| - $SWC_{0.6}$ | 13.4 (good) | 21 (good) | 19.2 (good) | 25.2 (good) | 20.4 (good) | 20.7 (good) | 18.6 (good) | 13.5 (good) | 18.6 (good) |
| - $SWC_{1.2}$ | 40.2 (good) | 42 (good) | 38.4 (good) | 50.4 (good) | 40.8 (good) | 41.4 (good) | 37.2 (good) | 27 (good) | 37.2 (good) |

SWC = smallest worthwhile percentage change, BS = Biering-Sorensen test, DEE = Dynamic Extensor Endurance test, SB-L = Side Bridge Left test, SB-R = Side Bridge Right test, BTC = Bench Trunk Curl-Up test.

analyses conducted revealed that all tests were sensitive enough to detect moderate to large changes in trunk muscle endurance.

Concerning the relative reliability, similar results (ICC > 0.75) were found in previous studies for the measures of BS,[28] DEE[27,28] and BTC.[25] The relative reliability scores of the measures obtained from the Ito and SB tests cannot be compared with the finding reported in previous studies because to the best of the authors' knowledge, this is the first study that has explored inter-session reliability in adolescents. The acceptable relative reliability results found in the present study for the measures of the 5 field-based tests selected might have been positively impacted by the large inter-individual variability observed in the trunk endurance scores (i.e. the 5 tests reported SDs larger than 20%). The heterogeneity documented for the participants' trunk endurance scores may be partially attributed to the large inter-individual difference (regarding level [magnitude of change], tempo [rate of change] and timing [onset of change]) in maturity status that is often found in adolescents within a given age group (up to 15 cm and 21 kg in the stature and body mass, respectively).[43,44]

The analysis of the presence of systematic bias between testing sessions revealed that this phenomenon only appeared in the trunk endurance measures obtained from the DEE and BTC tests, independent of the age group of the participants. Although for the DEE test, the presence of systematic bias was only found in boys when analysing bias by sex. These results were similar to the findings of Moya-Ramón et al.[25] who also found statistically significant inter-session differences in the endurance measure obtained through the BTC test in adolescents. The significant inter-session differences in BTC and DEE measures have been also confirmed by the 95% limits of agreement. For example, when an adolescent performed 60 repetitions on the BTC, on the retest he could perform as high as 60 + 9 = 69 repetitions or as low as 60 - 23.7 = 36.3 repetitions. An explanation for these significant inter-session differences may be based on the dynamic nature of both the DEE and BTC tests. For example, the execution of these two

tests requires participants to perform as many repetitions as possible with a certain cadence until exhaustion, which would entail the need of completing a pre-assessment familiarization session to learn their testing procedures when individuals who will be tested have little or no experience with them.[25,31,45] Given the time constraints in this study all participants were only allowed to freely practice the tests for ten minutes in each testing session. The main reason behind the implementation of this short and free familiarization protocol in each testing session was that all the field-based tests selected seem to present simple testing procedures. However, this short familiarization protocol may have not been sufficient for participants to learn how to maintain the specific cadence of movements required during the execution of the DEE and BTC tests.[25,31,42,46] Brotons-Gil et al. and Moya-Ramón et al. also carried out a similar familiarization protocol for their reliability study on dynamic trunk flexor endurance tests.[25,31] In this familiarization session, the subjects were informed of the test execution rules, but they did not perform the test, they only did 10 repetitions to familiarize themselves with the basic technique of the test. Although they performed the familiarization in a different session (1 week before), they also obtained significant inter-session differences.

The fact that all participants had limited or no experience with the trunk endurance field-based tests selected may explain the significant inter-session differences found for the measures from DEE and BTC and these differences might be attributed to a possible learning effect, which was consistent throughout the 3 age groups. Therefore, these findings suggest that in adolescents (12-18 years) with limited or no experience (independent of their age) with the DEE and BTC tests, the use of a longer familiarization protocol or the inclusion of an additional testing session to minimize the learning effects that were observed in the analyses of the inter-session differences.

Another factor that may have contributed to the systematic bias found in the endurance measures from the dynamic field-based tests is higher motivation in the second session by some

participants (information based on testers' comments). Although the trunk endurance scores (time and repetitions) achieved by participants in the five tests selected were not revealed by the testers until the end of the data collection phase, it might have been possible that some participants had mentally counted the number of repetitions completed during the two dynamic tests in the first session. Thus, in the second session, some participants exhibited significant motivation to achieve better scores in the dynamic tests than those obtained during the first session.[9,22,25,31,47] This circumstance was not as evident during the execution of the isometric tests because their final scores were determined in seconds rather than the number of repetitions.

The results of the present study also demonstrate that the precision of measurement of each field-based test could be categorized as poor. In particular, the $CV_{TE}$ of each field-based test ($CV_{TE}$ ranged from 11.3 to 24.4) exceeds the widely accepted 10% cut-off value to consider the magnitude of the measurement error as satisfactory, for both clinical and practice goals in healthy populations.[22,42] Only the $CV_{TE}$ of the endurance measure from the BTC ($CV_{TE}$ = 11.3) test approached this 10% cut-off score. Moya-Ramón et al.[25] also found poor $CV_{TE}$ scores for the BTC test in students (17.2%) aged 14-18. The precision of measurement of the rest of the field-based tests cannot be directly compared with previous studies as this is the first time that CVs have been calculated in an adolescent population. A plausible reason that may explain the poor $CV_{TE}$ scores of the trunk endurance measures found in this study may be also attributed to the fact that these physically demanding field-based tests are substantively influenced by motivation as no external/internal indicator of maximal effort is available. Consequently, these findings recommend that PE teachers convey to adolescents before carrying out the tests the vital importance that a maximal effort is required to achieve a true or real estimation of the trunk endurance capability. During the execution of each test, strong verbal encouragement is required for the adolescents to maintain the position required in the isometric tests for as long as possible, and to perform as many repetitions as possible in the dynamic tests. In turn, this

could also contribute to an improvement in the reliability scores of the trunk endurance measures recorded.

The $MDC_{95}$ value might be regarded as the minimum amount of change that needs to be observed, at either the group or individual level, for it to be considered a real or true change with a 95% level of certainty[48,49] (i.e. greater than the random measurement error). Thus, and according to the results of this study, changes higher than 31.4, 42.4, 44.5, 51.4, 54.6 and 67.5% for the BTC, DEE, BS, SB-L, SB-R and Ito tests (respectively) after an intervention may be considered as real or true with a 95% level of certainty.[50] The $MDC_{95}$ scores obtained for the BTC (31.4%) test could be acceptable for high-school-aged adolescents as previous studies have reported improvements of approximately 40% in trunk endurance after having completed a 6-week training program.[51–53] However, for the DEE, BS, SB-L, SB-R, and Ito tests these reported "improvements" of 40% in trunk endurance may not be true but simply measurement error. These findings are in line with the results of the sensitivity analyses that showed that these filed-based tests exhibited a good ability to detect moderate ($SWC_{0.6}$) to large ($SWC_{1.2}$) changes in their measures.

While the results of this study have provided information regarding the inter-session reliability of trunk muscle endurance measures obtained from five common field-based tests in an applied environment, limitations to the study must be acknowledged. Only high school-aged students were selected in this study and hence, the generalizability of the results to other populations (e.g.: adults) cannot be ascertained. However, the large sample size and the wide variety of trunk endurance field-based tests examined, in addition to the authenticity of the environment in which testing took place (high-school setting), are notable strengths of this study. Finally, for the dynamic field-based tests, the use of a longer familiarization protocol or the inclusion of an additional familiarization session or testing session might have minimized the learning effects that were observed in the analysis of the inter-session differences. Future applied studies

should explore the inter-session reliability of the trunk endurance measures obtained through filed-based tests using previous familiarization sessions in young populations. Similarly, whether the trunk endurance measures would be as reliable in a population of injured adolescents (e.g.: low back pain) must be considered, although these field-based tests are generally performed in healthy, uninjured populations. Besides, the lack of evidence regarding the inter-tester reliability of the measures obtained from the trunk extension endurance field-based tests in this cohort warrants further investigation.

To conclude, the findings from this study indicate that the trunk endurance measures obtained from five popular field-based tests (BTC, DEE, BS, SB-L, SB-R, and Ito) present poor inter-session reliability scores in high school-aged students (12-16 years). Even though all trunk endurance measures exhibited good relative reliability scores (ICC > 0.75), the precision of measurement of each field-based test was poor ($CV_{TE}$ > 10%). Therefore, we emphasize the importance of using both relative and absolute indices of reliability. Furthermore, inter-session systematic bias was observed in the dynamic trunk endurance field-based tests (BTC and DEE). The $MDC_{95}$ results indicate that changes higher than 31.4, 42.4, 44.5, 51.4, 54.6, and 67.5% for the BTC, DEE, BS, SB-L, SB-R, and Ito tests after an intervention may be considered as real or true with a 95% level of certainty. Only the BTC test obtained a $MDC_{95}$ low enough (31.4%) to be considered as acceptable for high school-aged students as previous studies have reported improvements of approximately 40% in trunk endurance after having completed a 6-week training program. The use of previous familiarization sessions before performing the tests (focus the attention on dynamic tests) and strongly encourage adolescents to do always their maximum effort in each test may help improve the reliability scores shown in this study.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**References**


1. Steffens D, Maher CG, Pereira LSM, Stevens ML, Oliveira VC, Chapple M, et al. Prevention of low back pain. A systematic review and meta-analysis. *JAMA Intern Med*. 2016;176(2):199-208. doi:10.1001/jamainternmed.2015.7431
2. Gomes-Neto M, Lopes JM, Conceição CS, Araujo A, Brasileiro A, Sousa C, et al. Stabilization exercise compared to general exercises or manual therapy for the management of low back pain: A systematic review and meta-analysis. *Phys Ther Sport*. 2017;23:136-142. doi:10.1016/j.ptsp.2016.08.004
3. De Blaiser C, Roosen P, Willems T, Danneels L, Bossche L Vanden, De Ridder R. Is core stability a risk factor for lower extremity injuries in an athletic population? A systematic review. *Phys Ther Sport*. 2018;30:48-56. doi:10.1016/j.ptsp.2017.08.076
4. Huxel Bliven KC, Anderson BE. Core Stability Training for Injury Prevention. *Sports Health*. 2013;5(6):514-522. doi:10.1177/1941738113481200
5. Biering-Sorensen F. Physical measurements as risk indicators for low-back trouble over a one-year period. *Spine (Phila Pa 1976)*. 1984;9(2):106-119.
6. Ito T, Shirado O, Suzuki H, Takahashi M, Kaneda K, Strax TE. Lumbar trunk muscle endurance testing: An inexpensive alternative to a machine for evaluation. *Arch Phys Med Rehabil*. 1996;77:75-79. doi:10.1016/S0003-9993(96)90224-5
7. McIntosh G, Wilson L, Affleck M, Hall H. Trunk and lower extremity muscle endurance: Normative data. *J Rehabil Outcomes Meas*. 1998;2(4):20-39.
8. McGill SM, Childs A, Liebenson C. Endurance times for low back stabilization exercises: Clinical targets for testing and training from a normal database. *Arch Phys Med Rehabil*. 1999;80(8):941-944. doi:10.1016/S0003-9993(99)90087-4
9. Geldhof E, Cardon G, De Bourdeaudhuij I, Danneels L, Coorevits P, Vanderstraeten G,

et al. Effects of back posture education on elementary schoolchildren's back function. *Eur Spine J*. 2007;16(6):829-839. doi:10.1007/s00586-006-0199-4

10. Bliss LS, Teeple P. Core stability: the centerpiece of any training program. *Curr Sports Med Rep*. 2005;4(3):179-183. doi:10.1007/s11932-005-0064-y
11. Luoto S, Heliövaara M, Hurri H, Alaranta H. Static back endurance low-back pain and the risk of low-back pain. *Clin Biomech*. 1995;10(6):323-324. doi:10.1016/0268-0033(95)00002-3
12. Knudson D, Johnston D. Validity and Reliability of a Bench Trunk-Curl Test of Abdominal Endurance. *J Strength Cond Res*. 1995;9(3):165-169.
13. Jetté M, Sidney K, Cicutti N. A critical analysis of sit-ups: a case for the partial curl-up as a test of abdominal muscular endurance. *CAHPER J*. 1984;51(1):4-9.
14. Robertson LD, Magnusdottir H. Evaluation of criteria associated with abdominal fitness testing. *Res Q Exerc Sport*. 1987;58(4):355-359. doi:10.1080/02701367.1987.10608112
15. Coorevits P, Danneels L, Cambier D, Ramon H, Vanderstraeten G. Assessment of the validity of the Biering-Sorensen test for measuring back muscle fatigue based on EMG median frequency characteristics of back and hip muscles. *J Electromyogr Kinesiol*. 2008;18(6):997-1005. doi:10.1016/j.jelekin.2007.10.012
16. Müller R, Strässle K, Wirth B. Isometric back muscle endurance: An EMG study on the criterion validity of the Ito test. *J Electromyogr Kinesiol*. 2010;20(5):845-850. doi:10.1016/j.jelekin.2010.04.004
17. Monfort-Pañego M, Vera-García FJ, Sánchez-Zuriaga D, Sarti-Martínez MÁ. Electromyographic Studies in Abdominal Exercises: A Literature Synthesis. *J Manipulative Physiol Ther*. 2009;32(3):232-244. doi:10.1016/j.jmpt.2009.02.007
18. Moreau CE, Green BN, Johnson CD, Moreau SR. Isometric back extension endurance

tests: A review of the literature. *J Manipulative Physiol Ther*. 2001;24(2):110-122. doi:10.1067/mmt.2001.112563

19. Juan-Recio C, Lopez-Vivancos A, Moya M, Sarabia JM, Vera-García FJ. Short-term effect of crunch exercise frequency on abdominal muscle endurance. *J Sports Med Phys Fitness*. 2015;55(4):280-289.

20. Juan-Recio C, Barbado D, López-Valenciano A, Vera-García FJ. Test de campo para valorar la resistencia de los músculos del tronco. *Apunt Educ Fis y Deport*. 2014;3(117):59-68. doi:10.5672/apunts.2014-0983.es.(2014/3).117.06

21. Juan-Recio C, López-Plaza D, Barbado D, García-Vaquero MP, Vera-García FJ. Reliability assessment and correlation analysis of 3 protocols to measure trunk muscle strength and endurance muscle strength and endurance. *J Sports Sci*. 2018;36(4):357-364. doi:10.1080/02640414.2017.1307439

22. Hopkins WG. Measures of Reliability in Sports Medicine and Science. *Sport Med*. 2000;30(5):375-381. doi:10.2165/00007256-200030050-00006

23. Matheson GJ. We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*. 2019;7(e6918). doi:10.7717/peerj.6918

24. Lloyd RS, Oliver JL, Faigenbaum AD, Myer GD, De Ste Croix MB. Chronological age vs. biological maturation: implications for exercise programming in youth. *J Strength Cond Res*. 2014;28(5):1454-1464. doi:10.1519/JSC.0000000000000391

25. Moya-Ramón M, Juan-Recio C, Lopez-Plaza D, Vera-Garcia FJ. Dynamic trunk muscle endurance profile in adolescents aged 14-18: Normative values for age and gender differences. *J Back Musculoskelet Rehabil*. 2018;31(1):155-162. doi:10.3233/BMR-169760

26. Martínez-Romero MT, Ayala F, De Ste Croix M, Vera-Garcia FJ, Sainz de Baranda P, Santonja-Medina F, et al. A Meta-Analysis of the Reliability of Four Field-Based

Trunk Extension Endurance Tests. *Int J Environ Res Public Health*. 2020;17(9). doi:10.3390/ijerph17093088

27. Hannibal III NS, Plowman SA, Looney MA, Brandenburg J. Reliability and Validity of Low Back Strength/Muscular Endurance Field Tests in Adolescents. *J Phys Act Health*. 2006;3(Suppl. 2):78-89. doi:https://doi.org/10.1123/jpah.3.s2.s78
28. Lin K-H, Huang Y-M, Tang W, Chang Y, Liu Y, Liu C. Correlation of static and dynamic trunk muscle endurance and bat swing velocity in high school aged baseball players. *Isokinet Exerc Sci*. 2013;21(2):113-119. doi:10.3233/IES-130486
29. Patterson P, Bennington J, de la Rosa T. Psychometric properties of child- and teacher-reported curl-up scores in children ages 10-12 years. *Res Q Exerc Sport*. 2001;72(2):117-124. doi:10.1080/02701367.2001.10608941
30. Boyer C, Tremblay M, Saunders T, McFarlane A, Borghese M, Lloyd M, et al. Feasibility, Validity, and Reliability of the Plank Isometric Hold as a Field-Based Assessment of Torso Muscular Endurance for Children 8-12 Years of Age. *Pediatr Exerc Sci*. 2013;25:407-422. doi:10.1123/pes.25.3.407
31. Brotons-Gil E, García-Vaquero M, Peco-González N, Vera-Garcia F. Flexion-Rotation Trunk Test To Assess Abdominal Muscle Endurance: Reliability, Learning Effect, and Sex Differences. *J Strength Cond Res*. 2013;27(6):1602-1608. doi:10.1519/JSC.0b013e31827124d9
32. De Blaiser C, De Ridder R, Willems T, Danneels L, Roosen P. Reliability and validity of trunk flexor and trunk extensor strength measurements using handheld dynamometry in a healthy athletic population. *Phys Ther Sport*. 2018;34:180-186. doi:10.1016/j.ptsp.2018.10.005
33. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive Statistics for Studies in Sports Medicine and Exercise Science. *Med Sci Sports Exerc*. 2009;41(1):3-12.

doi:10.1249/MSS.0b013e31818cb278

34. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64:96-106. doi:10.1016/j.jclinepi.2010.03.002

35. Impellizzeri FM, Marcora SM. Test validation in sport physiology: Lessons learned from clinimetrics. *Int J Sports Physiol Perform*. 2009;4:269-277. doi:10.1123/ijspp.4.2.269

36. Baumgartner TA, Chung H. Confidence Limits for Intraclass Reliability Coefficients. *Meas Phys Educ Exerc Sci*. 2001;5(3):179-188. doi:10.1207/S15327841MPEE0503

37. Lee MD, Wagenmakers EJ. *Bayesian Cognitive Modeling: A Practical Course*.; 2013. doi:10.1017/CBO9781139087759

38. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 2006;1:50-57. doi:10.1123/ijspp.1.1.50

39. Hinkle D, Wiersma W, Jurs S. *Applied Statistics for the Behavioral Sciences*. 5th ed. Houghton Mifflin; 2003. doi:10.2307/1164825

40. Hopkins WG. Spreadsheets for Analysis of Validity and Reliability. Sports Science. Published 2015. sportsci.org/2015/ValidRely.htm

41. Atkinson G, Nevill A. Statistical Methods for Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine. *Sport Med*. 1998;26(4):217-238. doi:10.2165/00007256-199826040-00002

42. Weir JP, Vincent WJ. *Statistics in Kinesiology*. 5th ed. Human Kinetics Publ Inc; 2020.

43. Evans K, Refshauge K, Adams R. Trunk muscle endurance tests: Reliability, and gender differences in athletes. *J Sci Med Sport*. 2007;10(6):447-455. doi:10.1016/j.jsams.2006.09.003

44. Radnor JM, Oliver JL, Waugh CM, Myer GD, Lloyd RS. The influence of maturity

status on muscle architecture in school-aged boys. *Pediatr Exerc Sci*. 2020;32(2):89-96. doi:10.1123/pes.2019-0201

45. De Blaiser C, De Ridder R, Willems T, Danneels L, Vanden Bossche L, Palmans T, et al. Evaluating abdominal core muscle fatigue: Assessment of the validity and reliability of the prone bridging test. *Scand J Med Sci Sport*. 2018;28(2):391-399. doi:10.1111/sms.12919
46. Strand SL, Hjelm J, Shoepe TC, Fajardo MA. Norms for an Isometric Muscle Endurance Test. *J Hum Kinet*. 2014;40(1):93-102. doi:10.2478/hukin-2014-0011
47. Narayan A, Steele-Johnson D. Relationships between prior experience of training, gender, goal orientation and training attitudes. *Int J Train Dev*. 2007;11(3):166-180. doi:10.1111/j.1468-2419.2007.00279.x
48. Beninato M, Portney LG. Applying concepts of responsiveness to patient management in neurologic physical therapy. *J Neurol Phys Ther*. 2011;35(2):75-81. doi:10.1097/NPT.0b013e318219308c
49. Portney LG. *Foundations of Clinical Research: Applications to Evidence-Based Practice*. 4th ed. FA Davis Company; 2020.
50. Furlan L, Sterr A. The applicability of standard error of measurement and minimal detectable change to motor learning research - A behavioral study. *Front Hum Neurosci*. 2018;12(95). doi:10.3389/fnhum.2018.00095
51. González-Gálvez N, Marcos-Pardo PJ, Carrasco-Poyatos M. Functional improvements after a pilates program in adolescents with a history of back pain: A randomised controlled trial. *Complement Ther Clin Pract*. 2019;35(January):1-7. doi:10.1016/j.ctcp.2019.01.006
52. Granacher U, Schellbach J, Klein K, Prieske O, Baeyens J, Muehlbauer T. Effects of core strength training using stable versus unstable surfaces on physical fitness in

adolescents: a randomized controlled trial. *BMC Sport Sci Med Rehabil*. 2014;6(40):1-11. doi:10.1186/2052-1847-6-40

53. Sandrey MA, Mitzel JG. Improvement in Dynamic Balance and Core Endurance after a 6-Week Core-Stability-Training Program in High School Track and Field Athletes. *J Sport Rehabil*. 2013;22(4):264-271. doi:10.1123/jsr.22.4.264