Test–Retest Reliability of Student-Administered Health-Related Fitness Tests in School Settings

Brendan T. O’Keeffe, Alan E. Donnelly, and Ciaran MacDonncha
University of Limerick

Purpose: To examine the test–retest reliability of student-administered (SA) health-related fitness tests in school settings and to compare indices of reliability with those taken by trained research-assistants. **Methods:** Participants ($n = 86$; age: 13.43 [0.33] y) were divided into 2 groups, SA ($n = 45$, girls = 26) or research-assistant administered (RA; $n = 41$, girls = 21). The SA group had their measures taken by 8 students (age: 15.59 [0.56] y, girls = 4), and the RA group had their measures taken by 8 research-assistants (age: 21.21 [1.38], girls = 5). Tests were administered twice by both groups, 1 week apart. Tests included body mass index, handgrip strength, standing broad jump, isometric plank hold, 90° push-up, 4 × 10-m shuttle run, back-saver sit and reach, and blood pressure. **Results:** Intraclass correlation coefficients for SA ($\geq .797$) and RA ($\geq .866$) groups were high, and the observed systematic error (Bland–Altman plot) between test 1 and test 2 was close to 0 for all tests. The coefficient of variation was less than 10% for all tests in the SA group, aside from the 90° push-up (24.3%). The SA group had a marginally lower combined mean coefficient of variation across all tests (6.5%) in comparison with the RA group (6.8%). **Conclusion:** This study demonstrates that, following familiarization training, SA health-related fitness tests in school-based physical education programs can be considered reliable.

Keywords: adolescent, youth, physical fitness, physical education

Physical fitness is a complex and multifaceted construct integrating a wide range of bodily functions including morphological, muscular, motor, cardiorespiratory, and metabolic (33). Physical fitness is composed of performance-related components and health-related components. In recent years, there has been a shift away from monitoring performance-related components of fitness to health-related components (36). Health-related physical fitness (HRPF) is made up of multiple components, including cardiorespiratory endurance (CRE); muscular fitness (muscular strength, local muscular endurance, and power); and body composition that have been identified as important markers of current and future health among children and adolescents (20,33,48). Higher levels of CRE are associated with reduced risk of future cardiometabolic-related diseases (46), potentially higher levels of academic achievement (4) and better mental health (33). In addition, positive changes to HRPF during adolescence can reduce the risk of negative health outcomes later in life (32,35). There is also mounting evidence that associates higher levels of muscular fitness in youth with lower levels of cardiovascular risk factors in young adulthood, independent of CRE and adiposity (17). In a systematic review of the health benefits of muscular fitness for children and adolescents, Smith et al (48) concluded that there was strong evidence of an inverse association between muscular fitness, central adiposity, and metabolic risk factors. The growing evidence base supporting the predictive capacity of physical fitness as a marker of current and future health has important implications for health promotion practices and has resulted in calls for the development of population wide monitoring of fitness (21,41).

Health-related physical fitness can be objectively and accurately measured in laboratory settings by qualified technicians using sophisticated instruments. However, as indicated by Espana-Romero et al (13), such tests are not feasible for administration at population level. Field-based tests provide a suitable alternative since they are time efficient, low in cost, and can be easily administered to a large number of people simultaneously (41), and there have been increasing calls for the development of simple, accurate, and inexpensive methods to measure fitness in youth (43). In any testing situation, it is important that the results are derived from high-quality measurement techniques. Reliability and validity are essential for meaningful interpretation and inference of results (29). Reliability refers to the reproducibility of a test result in repeated tests on the same individual under the same conditions. Extensive research has been conducted on the reliability of field-based measures of physical fitness. However, the majority of research to date has used an intratester or intertester reliability methodological design in which reliability is established by experienced and highly trained test administrators in standardized and controlled settings (13). Validity, described by Mahar and Rowe (26) as the most important concept in testing, refers to the ability of a test to reflect what it is designed to measure. A number of systematic reviews have recently been published identifying the criterion related validity of commonly administered field-based tests in youth, including the 20-m shuttle run (44), handgrip strength and standing broad jump (1), and body mass index (BMI) (36). Furthermore, as previously detailed, there is increasing evidence to support the predictive validity of body composition, CRE, and muscular fitness as powerful indicators of health in later life (40,41). Different models of HRPF test administration exist in secondary schools including: trained research-assistants visiting the school to collect HRPF data; the physical education teacher as the test coordinator and administrator; the physical education

The authors are with the Department of Physical Education and Sport Sciences, and the Health Research Institute, University of Limerick, Limerick, Ireland. O’Keeffe (brendan.okeeffe@ul.ie) is corresponding author.

teacher coordinates the test battery and fitness tests are administered by trained senior students; or finally, peer testing, where students measure each other's fitness. Given the greater degree of variability in secondary school settings, and the concurrent need for efficient test administration within a specified time period, maximizing validity and reliability represents a significant challenge.

Recently, a fitness test battery entitled ALPHA was developed specifically for use in school settings to facilitate monitoring fitness in a comparable way within the European Union (42). The ALPHA test battery was shown to be both valid and reliable when administered by physical education teachers in school settings (13). Other studies have also examined the reliability of teacher assessed measures of HRPF in school settings with positive results (38,53). Fitness tests are often administered in the form of a test battery, a set of 2 or more tests used to assess a component(s) of physical fitness. International examples currently in use in school settings include FitnessGram® (United States), CNPFT (China), ALPHA (European Union), Move! (Finland), GTO (Russia), SLOfit (Slovenia), and Netfit (Hungary). Several states in the United States and many countries including Japan, Finland, Slovenia, and Hungary have mandated monitoring physical fitness in school physical education programs (11,45,47). In such contexts, as noted in a recent review of HRPF monitoring practices in school-based physical education programs by O'Keeffe et al (30), it is often not feasible for one teacher to administer a test battery to a large group of students within the allocated time. The estimated time to perform the ALPHA priority test battery with 20 students was approximately 2 hours 30 minutes (42). Therefore, limited time and space present a significant challenge to test administrators, particularly in school contexts (12,34).

A student-administered (SA) format, where, following test protocol familiarization, students are responsible for the measurement of test items, could represent a feasible alternative (15,28). In a recent global review of youth fitness testing practices, Keating et al (22) found that of the 4 most prominent test batteries in use internationally, ALPHA (European Union), CNPFT (China), Fitnessgram (United States), and GTO (Russia), only Russia's GTO test battery supported a self-administration approach. Research has indicated that students are more in favor of a student-centered approach to fitness tests, as opposed to having their measurements taken by teachers or trained research-assistants (15,30). However, no research has been conducted to date on the reliability of SA measures of HRPF in school settings. Just as objective data are needed to answer key questions about the accuracy and repeatability of teacher-administered tests (26), so too are data on the reliability of SA measures. Therefore, the aim of this study was to determine if SA measures of HRPF in secondary school-based physical education programs can be considered reliable.

Methods

Participants

Research ethics approval for this study and the associated protocols was granted by the research ethics committee of the Faculty of Education and Health Sciences, University of Limerick, Ireland. The study was administered in a mixed-sex secondary school in the Midwest region of Ireland. All students (N=93) in year 1 of secondary education in the school were invited to participate. Informed consent to participate was received from the school principal, the participants, and their parents. The participation rate was 92.4% (N=86). The reliability of SA health-related fitness tests

was assessed using a 2-group design. Participants were assigned into a SA group (n=45; age: 13.44 [0.35] y; girls=26) or research assistant-administered (RA) group (n=41; age: 13.42 [0.32] y; girls=21) according to their timetabled physical education class. The RA group was included as a reference for expert-level reliability for this population.

The homogeneity of both the RA and SA groups was established through self-reported measures of physical activity (PACE+) (37) and the International Fitness Scale for youth (34) 1 week prior to commencing data collection. Age and gender were evenly distributed in both groups, and an independent sample *t* test indicated no significant differences between self-reported activity or fitness levels (*t* test, $P > .05$). Furthermore, following data collection, the authors explored homogeneity of variance in both groups by calculating the average of test 1 and test 2 (T1 and T2) for all fitness variables. Levene test indicated equal variances across all variables, $t_{84} = 3.5$, $P > .05$, with the exception of standing broad jump which was significantly higher in the RA group ($P = .37$), and an independent samples *t* test confirmed no statistically significant differences between mean scores of both groups (*t* test, $P > .05$), with the exception of handgrip strength which was significantly higher in the RA group (*t* test, $P = .01$).

Procedures

Both the SA and RA groups comprised of 8 test administrators. The cooperating physical education teacher selected 8 senior students (final 2 y of second-level education; age: 15.59 [0.56] y; girls=4) from the participating school to administer the test battery to the SA group. In addition, 8 research-assistants (≥ 2 y experience in field-based testing; age: 21.21 [1.38] y; girls=5) were recruited from the lead author's institution to administer the test battery to the RA group. Each administrator was responsible for one test item. A detailed standard operating procedure for each test item was designed for and read by both student and research-assistant test administrators 1 week before data collection started. Subsequently, test administrators from both groups participated in a 3-hour training workshop delivered by the lead author. During this workshop, each administrator was assigned one test and trained in the assigned test only. Student and research-assistant test administrators conducted several familiarization trials, and examples of correct and incorrect trials were demonstrated.

Reliability of each fitness test measure was established during a double period of physical education lasting 80 minutes. Tests were performed in a station format. Participants were provided with a minimum of 3 minutes rest between stations and 1 minute rest between trials. Participants performed the HRPF tests on 2 occasions (T1 and T2), on the same day at the same time, 1 week apart. Participant groupings and the order of test completion were the same on both test days for all participants. A period of 1 week for functional tests has been reported as sufficient to minimize learning effect, without introducing additional error due to maturation (25).

Measures

The tests included in this study were selected because they have high criterion validity (1,9,44), involve minimal equipment at low cost, and are feasible for administration in school settings (13,55). Test items included handgrip strength, standing broad jump, height, body mass, 4×10-m shuttle run. Four additional tests of physical fitness and health, commonly administered in

school-based HRPF test batteries and population health surveys, were included 90° push-up, isometric plank hold, back-saver sit and reach, and blood pressure (BP). The standard operating procedures used to administer each test are detailed below.

Anthropometry. Although not a direct measure of body composition, BMI represents a feasible alternative for use in school settings (36). BMI is a measure of weight for height and is calculated by dividing total body mass in kilograms by stature in meters squared. Body mass was measured to the nearest 0.1 kg using an electronic scale (range 0.05–200 kg; precision 0.05 kg; seca 875; seca, Birmingham, UK). Scales were calibrated using a known weight prior to testing. Stature was measured to the nearest 0.1 cm in the Frankfort plane with the participant standing upright (range 20 to 205 cm; precision 1 mm; seca 218; seca). During the anthropometric measurements, students wore light clothing and were barefoot. Both anthropometric measures were recorded twice. If a difference of >0.2 kg and/or 1 cm was recorded, participants were instructed to take a third measure. The mean of the 2 closest values was used for analysis.

Muscular Strength. Handgrip strength was measured using a digital hand dynamometer with adjustable grip (model: 5401; Takei Scientific Instruments Co, Ltd, Niigata, Japan). This dynamometer presents a high validity and reliability when calibrated with known weights (13). The grip span of the dynamometer was adjusted according to the hand size of the participant using an equation developed specifically for adolescents (43). Participants were instructed to squeeze the handle as hard as possible for 3 seconds, keeping the arm fully extended by the side of the body at all times. The test was performed twice and the maximum score for each hand was recorded in kilograms. The average of the scores achieved by left and right hands was used in the analysis. Lower body explosive strength was measured using the standing broad jump test. The participant stood on an Atrac® Jump Mat (Dewsbury, UK) behind the starting line and was instructed to push off vigorously and jump as far as possible. Following 3 submaximal practice trials, the test was repeated twice, and the best score was retained to the nearest centimeter as the distance between toes at take-off and heels at landing.

Muscular Endurance. Muscular endurance of the torso was measured using the isometric plank hold test. This test required participants to maintain a static prone position, with only forearms and toes touching the ground. Correct alignment required feet together with toes curled under the feet, elbows shoulder width distance apart, and forearms against the floor mat. Participants maintained eye contact with their hands, a neutral spine, and alignment from shoulders to ankles. The participant was given one 5-second practice trial, during which the test administrator instructed the participant into the correct position, followed by a brief period of rest. The timer started when the participant assumed the correct position. Participants were allowed to deviate from the correct position once and could continue the test if they immediately resumed the correct starting position. The test was terminated on the second deviation from the correct position or if the participant did not return to the correct position after the first warning. The score was recorded to the closest second using a stopwatch. Upper body muscular strength was measured using a 90° push-up test. The test was performed in line with the Fitnessgram® protocol as outlined by Welk and Meredith (55). Participants started in the push-up position, with their hands and toes touching the floor and arms shoulder width or slightly wider apart. Ensuring shoulder to ankle alignment, participants then lowered themselves toward the

ground until there was a 90° angle at the elbows, with upper arms parallel to the floor. A foam block was positioned under the participant to ensure a depth of 90° was reached before returning to the starting position. Push-ups were completed in time to a metronome set at 40 beats/min, with one complete push-up every 3 seconds. One form correction (eg, lowering of hips) was permitted. The test concluded on the second form correction or when the participant stopped due to fatigue.

Motor Component. Speed of movement and change of direction were assessed using the 4 × 10-m shuttle run test. The test was performed in line with the ALPHA test battery protocol (42). Two parallel lines were drawn on the floor 10 m apart. The participant ran as fast as possible from the starting line to the opposite line and returned to the starting line, crossing each line with both feet every time. This was performed twice, covering a distance of 40 m (4 × 10 m). Every time the participant crossed any of the lines, they were required to pick up (the first time) or exchange (second and third time) a sponge that was placed on either line prior to each trial. The timer was stopped when the participant crossed the finishing line with one foot. The time taken to complete the test was recorded to the nearest tenth of a second.

Flexibility. Hamstring and lumbar extensibility was measured using the back-saver sit and reach test following the protocols detailed by Welk and Meredith (55). Participants were asked to sit on the floor with legs fully outstretched. Feet, with shoes off, were placed with the soles flat against the test device (Cartwright Fitness® sit and reach box, Chester, UK). The knee of the participant's extended leg was held flat against the floor by the test administrator, hands aligned one over the other and palms facing down. Once the starting position was assumed, the participant reached forward along the measuring line as far as possible. To avoid a jerking action, participants were asked to hold their position at full extension for 3 seconds, before slowly returning to the starting position. The measurement device had a scale range of 70 cm and was marked in 0.5-cm intervals. The zero mark was 15 cm before the feet of the participant. The result was recorded to the nearest 0.5 cm, and the average of the highest scores achieved from the left and right side was used in the analysis.

Cardiovascular Health. Blood pressure was recorded using an Omron M6 (Matsusaka, Japan) automated oscillometric BP monitor. BP was measured by the test administrator according to the protocol outlined by the Centers for Disease Control and Prevention (8). Participants rested quietly for 3 to 5 minutes prior to the measurement. The participant was asked to sit all the way to the back of the chair so that the spine was straight. The left arm and back were fully supported, and legs were uncrossed with both feet flat on the floor. The left arm was unrestricted by clothing, with the palm of the hand turned upward and the elbow slightly flexed. The left arm was positioned so that the midpoint of the upper arm was at the approximate level of the heart.

Statistical Analyses

A Shapiro–Wilk test ($P > .05$) (39) and a visual inspection of their histograms showed that data were normally distributed, with the exception of SA BMI, handgrip strength, and isometric plank hold, and RA BMI and back-saver sit and reach ($P < .05$). Nonparametric alternatives were used to analyze these data where necessary. Sex specific effects on reliability were only found in SA systolic BP ($P = .02$, t test). Therefore, analyses were performed for both males and females together. A Pitman–Morgan test (14) indicated

homogeneity of variance for SA and RA groups across all measures between T1 and T2 ($P > .05$). In addition, the presence of heteroscedasticity was examined in line with the procedure as set out by Brehm et al (7). First, Bland–Altman plots were used to visually inspect the presence of heteroscedasticity by plotting the measurement differences (T2–T1) against the respective means. Following this, the degree of heteroscedasticity was then measured by calculating Kendall tau (τ_b) correlation between the absolute intertest difference and the corresponding means. When a positive correlation of $>.1$ was found, the data were denoted heteroscedastic. If heteroscedasticity was present, the data were transformed by logarithms to the base 10, if τ_b decreased, reliability was analyzed on the log-transformed scale (7).

The test–retest reliability of measures taken on both groups was explored using relative and absolute indices, and the results were then compared. Paired samples t tests (Wilcoxon signed-rank tests for nonparametric data) were used to determine systematic bias in mean values, intraclass correlation coefficient (ICC) was used to provide an estimate of rank order repeatability, and within-participant intertest variation was graphically illustrated using Bland–Altman plots (5). Mean and SD values for T1 and T2, as well as mean intertest differences (T2–T1), were calculated for both groups. The 95% limits of agreement for each physical fitness variable in both groups were calculated as the intertest mean difference ± 1.96 SD of the intertest differences. The Bland–Altman procedure considers the proportion between the magnitude of measurements and the error graphically, but not quantitatively. Therefore, the coefficient of variation (CV) between T1 and T2 was also calculated for each fitness test in both groups by dividing the SD by the mean and multiplying by 100 to get a percentage value. Atkinson and Nevill (3) note the advantage of using a dimensionless statistic such as the CV to facilitate comparison of reliability between different measurement tools or different groups using the same measurement tools, as was the case in this study. In an examination of the reliability of a battery of field-based fitness measures for adolescents, Lubans et al (24) suggested 20% variability was an acceptable degree of error. However, the decision as to what is acceptable agreement is a scientific judgment and one that statistics alone cannot answer (31), and thus, the threshold for

acceptable percentage error should be specific to the variable being measured (18). Therefore, a specific figure within which all tests might be considered reliable was not set. All calculations were performed using SPSS software for Windows (version 24.0; SPSS, Chicago, IL). For all analyses, the significance level was set at 5%.

Results

A total of 86 participants were assigned into a SA group ($n = 45$; age: 13.44 [0.35] y; girls = 26) or a RA group ($n = 41$; age: 13.42 [0.32] y; girls = 21). Mean and SD values for testing day 1 (T1) and testing day 2 (T2), as well as mean intertest differences (T2–T1), are reported in Table 1. It can be observed that intertest differences in the SA and RA groups were close to 0 for nearly all measurements. Highest mean intertest differences were observed in SA systolic BP (mean = 3.73 [7.4]) and RA standing broad jump (mean = –1.7 [0.1]).

Intraclass correlation coefficients, paired sample t tests, 95% limits of agreement (± 1.96 SD), and the CV for both RA and SA groups are reported in Table 2. ICC values for all tests in both SA (ICC $\geq .797$) and RA groups (ICC $\geq .866$) were high. An examination of systematic bias between T1 and T2 indicated no statistically significant intertest differences in either group ($P > .05$), aside from SA systolic BP (3.73 [7.35], $P = .002$) and RA 90° push-up (1.14 [1.98], $P = .003$). The CV was less than 10% for all tests in the SA group, aside from the 90° push-up (24.3%). The SA group had a lower mean CV in comparison with the RA group in 4 of the 9 tests administered, namely back-saver sit and reach, standing broad jump, isometric plank hold, and the 4 × 10-m shuttle run. Surprisingly, the SA group had a marginally lower combined mean CV across all tests (6.5%) in comparison with the RA group (6.8%).

Bland–Altman plots (Figures 1 and 2) were used to graphically show the reliability patterns in terms of systematic error (bias or mean intertest difference) and random error (95% limits of agreement) of the fitness tests studied. The systematic error, represented by the central line on the Bland–Altman plots, was close to 0 for all tests in both groups. A visual inspection of the plots indicated the presence of heteroscedasticity in SA handgrip strength and RA

Table 1 Test and Retest Measurements (Mean [SD]) of Student- and Research-Assistant Administered Groups

Variables	Student administered ($n = 45$; age: 13.44 [0.35]; girls = 26)			Research-assistant administered ($n = 41$; age: 13.42 [0.32]; girls = 21)		
	Test 1 Mean (SD)	Test 2 Mean (SD)	Intertest difference, Test 2 – Test 1 (SD)	Test 1 Mean (SD)	Test 2 Mean (SD)	Intertest difference, Test 2–Test 1 (SD)
BMI, kg/m ²	20.3 (3.6)	20.2 (3.7)	–0.05 (0.3)	20.4 (3.6)	20.4 (3.5)	–0.03 (0.2)
BSR, ^a cm	12.9 (5.9)	12.9 (5.9)	0.03 (1.5)	12.5 (8.6)	12.8 (8.6)	0.30 (2.4)
Systolic BP, mm Hg	101.5 (11.0)	104.9 (10.7)	3.73 (7.4)*	108.3 (13.0)	108.6 (12.0)	0.62 (16.2)
Diastolic BP, mm Hg	72.3 (7.2)	71.2 (8.6)	–1.15 (6.4)	69.4 (9.6)	69.6 (8.6)	0.21 (8.5)
Standing broad jump, cm	145.0 (20.6)	144.3 (20.8)	–0.70 (0.1)	158.3 (28.0)	156.6 (28.7)	–1.70 (0.1)
Handgrip, ^a kg	21.5 (4.1)	21.5 (4.5)	0.00 (1.1)	25.3 (5.4)	25.0 (5.3)	–0.30 (0.9)
90° push-up, repetitions	10.1 (6.6)	10.3 (6.3)	0.20 (2.8)	8.4 (6.5)	9.4 (6.4)	1.02 (2.0)*
Isometric plank hold, s	84.0 (39.6)	86.0 (41.1)	2.00 (16.1)	92.3 (43.7)	93.2 (43.1)	0.98 (16.3)
4 × 10-m shuttle run, s	12.2 (0.9)	12.2 (1.0)	0.01 (0.3)	11.6 (0.8)	11.5 (0.9)	–0.05 (0.4)

Abbreviations: BMI, body mass index; BP, blood pressure; BSR, back-saver sit and reach.

^aThe average of right and left side scores is shown in the table and was used for the analyses.

*Significant differences ($P < .05$) were found between trial 1 and trial 2, paired samples t test.

Table 2 Reliability Indices for Health and Physical Fitness Tests in Student- and Research-Assistant Administered Groups

Variables	Student-assistant measured (n = 45)				Research-assistant measured (n = 41)			
	ICC (95% CI)	P	LOA (± 1.96 SD)	CV, %	ICC (95% CI)	P	LOA (± 1.96 SD)	CV, %
BMI, kg/m ²	.998 (.996–.999)	.08 ^a	0.52 to –0.42	0.7	.999 (.998–1.00)	.93 ^{a,b}	0.41 to –0.41	0.6
BSR, cm	.984 (.971–.991)	.89	2.89 to –2.95	8.4	.980 (.962–.989)	.32 ^a	4.37 to –5.15	15.5
Systolic BP, mm Hg	.848 (.680–.922)	.002*	10.84 to –17.96	4.3	.900 (.812–.947)	.44	14.74 to –15.22	3.9
Diastolic BP, mm Hg	.797 (.632–.888)	.29	13.60 to –11.55	5.3	.866 (.747–.928)	.39	12.08 to –12.61	5.1
Standing broad jump, cm	.979 (.962–.988)	.45	0.12 to –0.11	2.0	.974 (.951–.986)	.23	0.19 to –0.16	3.2
Handgrip, kg	.984 (.970–.991)	.61 ^{a,b}	1.63 to –2.70	3.0	.992 (.985–.996)	.13	2.05 to –1.57	2.2
90° push-up, repetitions	.964 (.935–.980)	.42	4.35 to –5.02	24.3	.971 (.930–.986)	.003*	2.90 to –4.86	18.9
Isometric plank hold, s	.979 (.962–.989)	.46 ^a	24.55 to 20.51	9.8	.978 (.958–.989)	.77	24.23 to 25.86	10.4
4 × 10-m shuttle run, s	.985 (.973–.992)	.76	0.49 to –0.47	1.2	.943 (.891–.970)	.42	0.83 to –0.73	1.8

Abbreviations: BMI, body mass index; BP, blood pressure; BSR, back-saver sit and reach; CI, confidence interval; CV, coefficient of variation; ICC, intraclass correlation coefficient; LOA, limits of agreement (mean differences ± 1.96 SD).

^aWilcoxon signed-rank test used for nonnormally distributed data. ^bIf heteroscedasticity was present, the data were transformed by logarithms to the base 10.

*Significant differences ($P < .05$) were found between trial 1 and trial 2, paired samples t test.

BMI. This was confirmed by Kendall tau (τ_b) correlation coefficient values of .264 and .259 for SA handgrip and RA BMI, respectively. Log transformations to the base 10 did not remove the heteroscedasticity for either measure ($P = .01$, SA handgrip strength; $P = .02$, RA BMI). Heteroscedasticity was not observed for any other measure in both groups ($P \geq .05$).

Discussion

The results from this study offer insights about the quality of SA fitness tests in school settings. The practicality and feasibility of field-based tests for administration in school settings are crucial; however, students, parents, and policymakers need to have confidence in the validity and reliability of the data gathered. With the aim of making data gathered as close to the reality of a school context as possible, this study was performed during timetabled physical education lessons, and both student and research-assistant test administrators received the same test protocol administration training 1 week in advance of testing. The main findings suggest that, following training on test administration protocols, SA fitness tests including BMI, back-saver sit and reach, standing broad jump, handgrip strength, isometric plank hold, and the 4 × 10-m shuttle run can be considered a reliable alternative to RA tests. Further research is needed to confirm the reliability of the 90° push-up and BP tests in light of the large variations found between T1 and T2 in both SA and RA groups.

This is the first study of its kind to examine the intrarater reliability of SA physical fitness tests and to compare these reliability indices with those taken by research-assistants. Surprisingly, the SA group had a marginally lower combined mean CV value (6.5%) than the RA group (6.8%) across all tests, potential reasons for which are explored later in the discussion. Aside from SA BP, in which girls varied significantly more than boys ($P \leq .05$), no sex specific differences were observed between T1 and T2 in either group. However, it should be noted that despite reaching statistical significance, mean differences in SA systolic BP recordings between T1 and T2 among girls were relatively small (< 3 mm Hg). Relative reliability indices were good, with high ICCs across all tests in both the SA (ICC $\geq .797$) and RA (ICC $\geq .866$) groups. These values compare favorably to similar studies including Lubans et al (24) who reported ICCs of $\geq .785$ in an examination

of the reliability of commonly administered field-based fitness measures in adolescents. Systematic bias was only observed in 2 tests, SA 90° push-up, and RA systolic BP; however, the difference in mean values was small for both tests, as detailed in Table 1. A significant association between the magnitude of the measure and the difference between test and retest values (heteroscedasticity) was only observed in SA handgrip strength and RA BMI. In both cases, higher values produced significantly more variability in test–retest scores ($P \leq .05$). All other tests analyzed were homoscedastic. In an examination of the reliability of fitness tests administered by teachers in schools, Espana-Romero et al (13) did not find the presence of heteroscedasticity in any fitness variable they analyzed. In a similar study, Ramírez-Vélez et al (38) did note the presence of heteroscedasticity for the back-saver sit and reach; however, this was not observed in this study.

Although not strictly a measure of body composition, the Institute of Medicine recommends BMI as the most appropriate anthropometric measure for use in schools (36). The reliability of BMI measures taken in school-based physical education programs has been shown to be higher when compared with other measures of body composition (13). Relative and absolute reliability indices were very high for measures of BMI in both the SA (ICC $\geq .797$; CV = 0.7%) and RA (ICC $\geq .866$; CV = 0.6%) groups. Espana-Romero et al (13) also reported measurement error values of $< 1\%$ for BMI in their analysis of the reliability of teacher-administered fitness tests in school settings. High interrater reliability of BMI measures taken by a school nurse and trained research staff among boys and girls aged 5–12 years has also been reported elsewhere (51). The reliability of BMI measures may be enhanced when 2 or more measures are obtained and averaged (54), as was the protocol used in this study. While some scholars have highlighted the benefits of systematic monitoring of anthropometric measures in school settings (52), physical education teachers need to be mindful of the influence of peer pressure, body image concerns, and elevated levels of anxiety on self-efficacy in fitness test performance (23). The testing station format of delivery, as outlined in the “Methods” section, allowed student participants to perform tests in small groups, potentially alleviating the prominence of body image concerns that may be present when fitness testing in larger groups. Regardless, special care needs to be given to preparing student testers in appropriately administering

Figure 1 — Bland–Altman plots for student-administered BMI, back-saver sit and reach, standing broad jump, handgrip strength, 90° push-up, isometric plank hold, diastolic blood pressure, and 4 × 10-m shuttle run. The central line represents the mean differences between the T2 and the T1; the upper and lower black lines represent the upper and lower 95% limits of agreement (means differences $\pm$ 1.96 SD of the differences), respectively. BMI indicates body mass index; T1, first test; T2, second test.

anthropometric tests to their peers in school settings, particular in coeducational or mixed-sex schools, and such tests should only be used in contexts in which the physical education teacher deems it appropriate.

Reliability patterns for tests measuring musculoskeletal fitness varied considerably. CV values for the 90° push-up were very

high in both the SA group (24.3%) and RA group (18.9%), despite having ICC high values of ≥ 0.964 (SA group) and ≥ 0.971 (RA group). The wide range of scores observed in each group could explain this discrepancy as reported by Atkinson and Nevill (3), further emphasizing the importance of using both relative and absolute indices of reliability. Lubans et al (24) and Morrow et al (29)

Figure 2 — Bland–Altman plots for research assistant–administered BMI, back-saver sit and reach, standing broad jump, handgrip strength, isometric plank hold, 4 × 10-m shuttle run test, diastolic blood pressure, and systolic blood pressure. The central line represents the mean differences between the T2 and the T1; the upper and lower black lines represent the upper and lower 95% limits of agreement (means differences ± 1.96 SD of the differences), respectively. BMI indicates body mass index; T1, first test; T2, second test.

found similarly poor reliability for the 90° push-up test, suggesting that the current protocols used to administer this test are inadequate to produce sufficiently reliable results. Interestingly, a significant intertest difference was found for RA 90° push-up test ($P = .003$)

that was not observed in the SA group. This suggests that reliability declined with experience; however, it has been reported that more experienced testers may not count as many repetitions that meet the criteria for a full repetition (elbow flexion of 90°), hence the greater

degree of variability (27). A similar finding of reduced reliability with tester experience was observed by Morrow et al (29) for the trunk-lift test. CV values for the isometric plank hold tests were similar, 9.8% and 10.4% for SA and RA groups, respectively. Acceptable reliability of the isometric plank hold test among children aged 8–12 years has been found elsewhere (6); however, this study only reported ICC values, the limitations of which have previously been outlined. Test–retest reliability for the handgrip strength test and standing broad jump was very high in the SA and RA groups. This corroborates the findings of previous studies that examined the reliability of both test items delivered by physical education specialists in school settings (13,31,38). Espana-Romero et al (13) reported marginally lower percentage error values of 2.3% for handgrip strength and 6.3% for standing broad jump in comparison with those observed in this study (Table 2). In a systematic review on the reliability of field-based tests in youth, Artero et al (2) also indicated that neither learning nor fatigue effects were found for either the standing broad jump or handgrip strength tests in the studies they identified. Therefore, student assessed measures of muscular fitness including handgrip strength, standing broad jump, and isometric plank hold can be considered reliable for administration in school settings.

The back-saver sit and reach test produced high relative reliability scores in both the SA (ICC = .984) and RA (ICC = .980) groups. Hartman and Looney (19) found similarly high ICC levels of .975 for the same test in a study involving 87 boys and 62 girls aged 6–12 years. Notably, the SA groups' mean CV value of 8.4% for the back-saver sit and reach was almost half that of the RA group (15.5%). As outlined previously for the 90° push-up test, variability in terms of what a test administrator considers an adequate trial could be the source of this greater discrepancy among the more experienced RA testers. In line with existing research (13,38), the 4 × 10-m shuttle had very good relative and absolute reliability indices in both groups. In an analysis of reliability of teacher-administered HRPf tests among Columbian children and adolescents, Ramírez-Vélez et al (38) found similarly low intertest differences in the 4 × 10-m shuttle run of less than three tenths of a second. The 4 × 10-m shuttle run test thus represents a very practical and reliable approach to assessing components of motor fitness, including speed and change of direction, among youth. Automated BP recordings, although within a moderate range, had the lowest ICC values for both student- and research-administered groups when compared with the other test items (RA: systolic ICC = .677, diastolic ICC = .761; SA: systolic ICC = .742, diastolic ICC = .668). Despite this, CV values for BP were ≤5.3% for both groups. Previous studies have highlighted multiple advantages of automated BP recordings including practicality and reliability (10), as well as ease of administration by nonexpert populations (16,50). To optimize validity of a BP recording, a 24-hour ambulatory BP monitoring protocol is recommended (49), which is not feasible in a school context. Further research is needed to determine the reliability and validity of field-based BP measurements administered by nonexpert populations.

This study had some limitations which should be noted. Given the relatively small sample size and tight age range of participants drawn from only one school, our findings cannot be generalized to all field-based testing settings at this time. A larger sample size, involving a more diverse age range, and the inclusion of an additional trial, could have improved the precision of the reliability estimates, while also allowing for a more detailed examination of results by age group. However, the 2-group study design (SA and RA) and the wide variety fitness tests examined, in addition to the authenticity of the environment in which testing took place, are

notable strengths of this study. Although previous research has promoted the concept of peer fitness testing in a school context, this study is the first of its kind to examine and confirm the reliability of this approach. The findings indicate that, with adequate training on test administration protocols, a student-led approach represents a feasible and reliable alternative to a physical education teacher or RA fitness tests in school settings.

Conclusion

The aim of this study was to examine the intrarater reliability of SA HRPf tests in a school context. Student-centered approaches to fitness testing in school contexts have long been recommended; however, this is the first study to examine the reliability of student measured fitness tests. Although no testing situation can be perfect, particularly in a field-based or school context, this study presents various steps that can be taken to minimize potential sources of error and optimize reliability, while simultaneously contributing to student learning. Overall, SA reliability indices were very positive, and rather surprisingly, the SA group had a marginally lower combined CV across all tests in comparison with the RA group. The results suggest that, following training on test administration protocols, student-assessed measures of anthropometric (BMI); muscular fitness (handgrip strength, standing broad jump, and isometric plank hold); and performance-related fitness (4 × 10-m shuttle run) can be considered a reliable approach to administering HRPf tests in school settings. A SA approach to fitness testing in school settings represents an accessible and feasible mechanism for gathering data on key indicators of adolescent health. However, any measurement of physical fitness in a school context should be delivered with a strong educational emphasis, and not conducted solely for the purpose of gathering data.

Acknowledgment

The study received support funding from the Government of Ireland, Irish Research Council Postgraduate Scholarship Scheme.

References

1. Artero E, España-Romero V, Castro-Piñero J, Ortega FB, Sjöström M, Suni J, Ruiz JR. Criterion-related validity of field-based muscular fitness tests in youth. *J Sports Med Phys Fitness*. 2012;52(3):263–72. PubMed ID: 22648464
2. Artero EG, Espana-Romero V, Castro-Piñero J, Ortega FB, Suni J, Castillo-Garzon MJ, Ruiz JR. Reliability of field-based fitness tests in youth. *Int J Sports Med*. 2011;32(3):159–69. PubMed ID: 21165805 doi:10.1055/s-0030-1268488
3. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26(4):217–38. PubMed ID: 9820922 doi:10.2165/00007256-199826040-00002
4. Bezold CP, Konty KJ, Day SE, et al. The effects of changes in physical fitness on academic performance among New York City youth. *J Adolesc Health*. 2014;55(6):774–81. PubMed ID: 25088395 doi:10.1016/j.jadohealth.2014.06.006
5. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*. 1995;346(8982):1085–7. PubMed ID: 7564793 doi:10.1016/S0140-6736(95)91748-9
6. Boyer C, Tremblay M, Saunders T, McFarlane A, Borghese M, Lloyd M, Longmuir P. 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(3):407–22. PubMed ID: [23877226](#) doi:[10.1123/pes.25.3.407](#)
7. Brehm MA, Scholtes VA, Dallmeijer AJ, Twisk JW, Harlaar J. The importance of addressing heteroscedasticity in the reliability analysis of ratio-scaled variables: an example based on walking energy-cost measurements. *Dev Med Child Neurol*. 2012;54(3):267–73. PubMed ID: [22150364](#) doi:[10.1111/j.1469-8749.2011.04164.x](#)
 8. Centers for Disease Control and Prevention. *National Center for Health Statistics (NCHS) Laboratory Procedures Manual. National Health and Nutrition Examination Survey Questionnaire (Blood Pressure Protocol)*. Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention; 2008. http://www.cdc.gov/nchs/data/nhanes/nhanes_09_10/lab.pdf. Accessed June 22, 2015.
 9. Chen Z, Wang X, Wang Z, et al. Assessing the validity of oscillometric device for blood pressure measurement in a large population-based epidemiologic study. *J Am Soc Hypertens*. 2017;11(11):730–6.e4. PubMed ID: [29032943](#) doi:[10.1016/j.jash.2017.09.004](#)
 10. Christofaro DGD, Casonatto J, Polito MD, et al. Evaluation of the Omron MX3 Plus monitor for blood pressure measurement in adolescents. *Eur J Pediatr*. 2009;168(11):1349. PubMed ID: [19221789](#) doi:[10.1007/s00431-009-0936-x](#)
 11. Csányi T, Finn KJ, Welk GJ, et al. Overview of the Hungarian National Youth fitness study. *Res Q Exerc Sport*. 2015;86(suppl 1):S3–12. doi:[10.1080/02701367.2015.1042823](#)
 12. De Moraes ACF, Vilanova-Campelo RC, Torres-Leal FL, Carvalho HB. Is self-reported physical fitness useful for estimating fitness levels in children and adolescents? A reliability and validity study. *Medicina*. 2019;55(6):286. doi:[10.3390/medicina55060286](#)
 13. Espana-Romero V, Artero EG, Jimenez-Pavon D, et al. Assessing health-related fitness tests in the school setting: reliability, feasibility and safety; the ALPHA Study. *Int J Sports Med*. 2010;31(7):490–7. PubMed ID: [20432194](#) doi:[10.1055/s-0030-1251990](#)
 14. Gardner RC. *Psychological Statistics Using SPSS for Windows*. Upper Saddle River, NJ: Prentice Hall; 2001.
 15. Graser SV, Sampson BB, Pennington TR, Prusak KA. Children's perceptions of fitness self-testing, the purpose of fitness testing, and personal health. *Phys Educ*. 2011;68(4):175–87.
 16. Graves JW, Althaf MM. Utility of ambulatory blood pressure monitoring in children and adolescents. *Pediatr Nephrol*. 2006;21(11):1640–52. PubMed ID: [16823576](#) doi:[10.1007/s00467-006-0175-6](#)
 17. Grøntved A, Ried-Larsen M, Møller NC, Kristensen PL, Froberg K, Brage S, Andersen LB. Muscle strength in youth and cardiovascular risk in young adulthood (the European Youth Heart Study). *Br J Sports Med*. 2015;49(2):90–4. doi:[10.1136/bjsports-2012-091907](#)
 18. Hanneman SK. Design, analysis and interpretation of method-comparison studies. *AACN Adv Crit Care*. 2008;19(2):223. PubMed ID: [18560291](#)
 19. Hartman JG, Looney M. Norm-referenced and criterion-referenced reliability and validity of the back-saver sit-and-reach. *Meas Phys Educ Exerc Sci*. 2003;7(2):71–87. doi:[10.1207/S15327841MPEE0702_2](#)
 20. Hurtig-Wennlöf A, Ruiz JR, Harro M, Sjöström M. Cardiorespiratory fitness relates more strongly than physical activity to cardiovascular disease risk factors in healthy children and adolescents: the European Youth Heart Study. *Eur J Cardiovasc Prev Rehabil*. 2007;14(4):575–81. doi:[10.1097/HJR.0b013e32808c67e3](#)
 21. Kaminsky LA, Arena R, Beckie TM, et al. The importance of cardiorespiratory fitness in the United States: the need for a national registry: a policy statement from the American Heart Association. *Circulation*. 2013;127(5):652–62. PubMed ID: [23295916](#) doi:[10.1161/CIR.0b013e31827ee100](#)
 22. Keating XD, Smolianov P, Liu X, Castro-Piñero J, Smith J. Youth fitness testing practices: global trends and new development. *Sport J*. 2018.
 23. Lodewyk KR, Sullivan P. Associations between anxiety, self-efficacy, and outcomes by gender and body size dissatisfaction during fitness in high school physical education. *Phys Educ Sport Pedagogy*. 2016;21(6):603–15. doi:[10.1080/17408989.2015.1095869](#)
 24. Lubans DR, Morgan P, Callister R, Plotnikoff RC, Eather N, Riley N, Smith CJ. Test-retest reliability of a battery of field-based health-related fitness measures for adolescents. *J Sports Sci*. 2011;29(7):685–93. PubMed ID: [21391082](#) doi:[10.1080/02640414.2010.551215](#)
 25. Lubans DR, Smith JJ, Harries SK, Barnett LM, Faigenbaum AD. Development, test-retest reliability, and construct validity of the resistance training skills battery. *J Strength Cond Res*. 2014;28(5):1373–80. PubMed ID: [24755868](#) doi:[10.1519/JSC.0b013e31829b5527](#)
 26. Mahar MT, Rowe DA. Practical guidelines for valid and reliable youth fitness testing. *Meas Phys Educ Exerc Sci*. 2008;12(3):126–45. doi:[10.1080/10913670802216106](#)
 27. McManis BG, Baumgartner TA, Wuest DA. Objectivity and reliability of the 90 push-up test. *Meas Phys Educ Exerc Sci*. 2000;4(1):57–67. doi:[10.1207/S15327841Mpee0401_6](#)
 28. Morrow JR Jr, Ede A. Research quarterly for exercise and sport lecture statewide physical fitness testing: a BIG Waist or a BIG Waste? *Res Q Exerc Sport*. 2009;80(4):696–701. PubMed ID: [20025110](#)
 29. Morrow JR Jr, Martin SB, Jackson AW. Reliability and validity of the FITNESSGRAM: quality of teacher-collected health-related fitness surveillance data. *Res Q Exerc Sport*. 2010;81(3)(suppl):S24–30. PubMed ID: [21049835](#) doi:[10.1080/02701367.2010.10599691](#)
 30. O'Keeffe BT, MacDonncha C, Ng K, Donnelly AE. Health-related fitness monitoring practices in secondary school-based physical education programs. *J Teach Phys Educ*. Epub ahead of print. doi:[10.1123/jtpe.2018-0336](#)
 31. Ortega FB, Artero EG, Ruiz JR, et al. Reliability of health-related physical fitness tests in European adolescents. The Helena Study. *Int J Obes*. 2008;32(suppl 5):S49–57. doi:[10.1038/ijo.2008.183](#)
 32. Ortega FB, Labayen I, Ruiz JR, et al. Improvements in fitness reduce the risk of becoming overweight across puberty. *Med Sci Sports Exerc*. 2011;43(10):1891–7. PubMed ID: [21407124](#)
 33. Ortega FB, Ruiz JR, Castillo MJ, Sjöström M. Physical fitness in childhood and adolescence: a powerful marker of health. *Int J Obes*. 2008;32(1):1–11. doi:[10.1038/sj.ijo.0803774](#)
 34. Ortega FB, Ruiz JR, Espana-Romero V, et al. The International Fitness Scale (IFIS): usefulness of self-reported fitness in youth. *Int J Epidemiol*. 2011;40(3):701–11. PubMed ID: [21441238](#) doi:[10.1093/ije/dyr039](#)
 35. Ortega FB, Silventoinen K, Tynelius P, Rasmussen F. Muscular strength in male adolescents and premature death: cohort study of one million participants. *BMJ*. 2012;345:e7279. PubMed ID: [23169869](#) doi:[10.1136/bmj.e7279](#)
 36. Pillsbury L, Oria M, Pate RR. *Fitness Measures and Health Outcomes in Youth*. Washington, DC: National Academies Press; 2013.
 37. Prochaska JJ, Sallis JF, Long B. A physical activity screening measure for use with adolescents in primary care. *Arch Pediatr Adolesc Med*. 2001;155(5):554–9. PubMed ID: [11343497](#) doi:[10.1001/archpedi.155.5.554](#)
 38. Ramírez-Vélez R, Rodrigues-Bezerra D, Correa-Bautista JE, Izquierdo M, Lobelo F. Reliability of health-related physical fitness tests among Colombian children and adolescents: The FUPRECOL Study. *PLoS One*. 2015;10(10):e0140875. doi:[10.1371/journal.pone.0140875](#)

39. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Anal.* 2011;2(1):21–33.
40. Rodrigues LP, Leitaó R, Lopes VP. Physical fitness predicts adiposity longitudinal changes over childhood and adolescence. *J Sci Med Sport.* 2013;16(2):118–23. PubMed ID: [22824312](#) doi:[10.1016/j.jsams.2012.06.008](#)
41. Ruiz JR, Castro-Piñero J, Artero EG, Ortega FB, Sjöström M, Suni J, Castillo MJ. Predictive validity of health-related fitness in youth: a systematic review. *Br J Sports Med.* 2009;43(12):909–23. PubMed ID: [19158130](#) doi:[10.1136/bjsm.2008.056499](#)
42. Ruiz JR, Castro-Piñero J, España-Romero V, et al. Field-based fitness assessment in young people: the ALPHA health-related fitness test battery for children and adolescents. *Br J Sports Med.* 2011; 45(6):518–24. PubMed ID: [20961915](#) doi:[10.1136/bjsm.2010.075341](#)
43. Ruiz JR, Ortega FB, Gutierrez A, Meusel D, Sjöström M, Castillo MJ. Health-related fitness assessment in childhood and adolescence: a European approach based on the AVENA, EYHS and HELENA studies. *J Public Health.* 2006;14(5):269–77. doi:[10.1007/s10389-006-0059-z](#)
44. Ruiz JR, Silva G, Oliveira N, Ribeiro JC, Oliveira JF, Mota J. Criterion-related validity of the 20-m shuttle run test in youths aged 13–19 years. *J Sports Sci.* 2009;27(9):899–906. PubMed ID: [19629839](#) doi:[10.1080/02640410902902835](#)
45. Salin K, Huhtiniemi M. Physical Education in Finland after curriculum reform 2016. In: Popović S, Antala B, Bjelica D, Gardašević J, (Eds.). *Physical Education in Secondary School: Researches, Best Practices, Situation.* Podgorica, Montenegro: Montenegro Faculty of Sport and Physical Education of University of Montenegro, Montenegrin Sports Academy and FIEP; 2018:329–34.
46. Schmidt M, Magnussen C, Rees E, Dwyer T, Venn A. Childhood fitness reduces the long-term cardiometabolic risks associated with childhood obesity. *Int J Obes.* 2016;40(7):1134–40. doi:[10.1038/ijo.2016.61](#)
47. Shephard RJ. *A History of Health & Fitness: Implications for Policy Today.* New York, NY: Springer; 2018.
48. Smith J, Eather N, Morgan P, Plotnikoff R, Faigenbaum A, Lubans D. The health benefits of muscular fitness for children and adolescents: a systematic review and meta-analysis. *Sports Med.* 2014;44(9): 1209–23. PubMed ID: [24788950](#) doi:[10.1007/s40279-014-0196-4](#)
49. Soergel M, Kirschstein M, Busch C, et al. Oscillometric twenty-four-hour ambulatory blood pressure values in healthy children and adolescents: a multicenter trial including 1141 subjects. *J Pediatr.* 1997;130(2):178–84. PubMed ID: [9042117](#) doi:[10.1016/S0022-3476\(97\)70340-8](#)
50. Stergiou GS, Karpettas N, Kapoyiannis A, Stefanidis CJ, Vazeou A. Home blood pressure monitoring in children and adolescents: a systematic review. *J Hypertens.* 2009;27(10):1941–7. PubMed ID: [19542894](#) doi:[10.1097/HJH.0b013e32832ea93e](#)
51. Stoddard SA, Kubik MY, Skay C. Is school-based height and weight screening of elementary students private and reliable? *J Sch Nurs.* 2008;24(1):43–8. PubMed ID: [18220455](#) doi:[10.1177/10598405080240010701](#)
52. Thompson HR, Linchey JK, King B, Himes JH, Madsen KA. Accuracy of school staff-measured height and weight used for body mass index screening and reporting. *J Sch Health.* 2019; 89(8):629–35. PubMed ID: [31140199](#) doi:[10.1111/josh.12788](#)
53. Vanhelst J, Béghin L, Fardy PS, Ulmer Z, Czaplicki G. Reliability of health-related physical fitness tests in adolescents: the MOVE Program. *Clin Physiol Funct Imaging.* 2016;36(2):106–11. PubMed ID: [25319253](#) doi:[10.1111/cpf.12202](#)
54. Vegelin A, Brukx L, Waelkens J, Van den Broeck J. Influence of knowledge, training and experience of observers on the reliability of anthropometric measurements in children. *Ann Hum Biol.* 2003;30(1):65–79. PubMed ID: [12519655](#) doi:[10.1080/03014460210162019](#)
55. Welk G, Meredith MD. *Fitnessgram and Activitygram Test Administration Manual-Updated.* 4th ed. Champaign, IL: Human Kinetics; 2010.