Test-Retest Reliability of a Battery of Field-Based Health-Related Fitness Measures for Adolescents

Running head: Adolescent Fitness Tests

David R. Lubans[1]*, Philip J. Morgan[1], Robin Callister[2], Ronald C. Plotnikoff[1], Narelle Eather[1], Nicholas Riley[1] and Chris J. Smith[2]

[1]School of Education, University of Newcastle, Callaghan Campus, AUSTRALIA;

[2]School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan Campus, AUSTRALIA;

* Corresponding author
David Lubans, PhD
University of Newcastle
School of Education
Callaghan Campus
NSW 2308
Australia
Email: David.Lubans@newcastle.edu.au
Telephone: +61 2 49212049
Fax: +61 2 49217407

Key words: Physical fitness; physical activity; resistance training; strength training; muscular fitness; body composition

Abstract

The primary aim of this study was to determine the test-retest reliability of existing tests of health-related fitness (HRF). Participants (mean age = 14.8 ±0.4 years) were 42 boys and 26 girls who completed the study assessments on two occasions separated by one week. The following tests were conducted: bio-electrical impedance analysis (BIA) to calculate % body fat, leg dynamometer, 90 degree push-up, 7-stage sit-up and wall squat tests. Intraclass correlation (ICC), paired samples t-tests, and typical error expressed as a coefficient of variation (CV) were calculated. The body fat % ICC values were similar for boys (ICC = 0.95) and girls (ICC = 0.93), but the CV was considerably higher for boys (22.2% versus 12.2%). The boys' CV values for the tests of muscular fitness ranged from 9.0% for the leg dynamometer test to 26.5% for the timed wall squat test. The girls' CV values ranged from 17.1% for the sit-up test to 21.4% for the push-up test. While the BIA machine produced reliable estimates of % body fat, the tests of muscular fitness resulted in high levels of systematic error suggesting that these measures may require an extensive familiarisation phase before the results can be considered reliable.

Test-Retest Reliability of a Battery of Field-Based Health-Related Fitness Measures for Adolescents

While it remains contentious whether physical activity or physical fitness is more important to the health of young people (Blair, Cheng, & Holder, 2001; Boreham & Riddoch, 2001), recent studies have found that measures of physical fitness are more strongly associated with indicators of health among youth than time spent in physical activity (Ekelund, et al., 2001; Rizzo, Ruiz, Hurtig-Wennlof, Ortega, & Sjostrom, 2007). Furthermore, a recent systematic review concluded that improvements in muscular strength from childhood to adolescence are negatively associated with changes in adiposity and a healthier body composition during adolescence is predictive of an improved cardiovascular profile later in life (Ruiz, et al., 2009). These findings have important implications for health promotion and justify strategies by health professionals to promote physical activity of sufficient intensity to increase fitness, especially muscular fitness.

Resistance training (RT) is exercise designed specifically to increase muscular fitness through increased workload demand and may include the use of free weights, machine weights, elastic-tubing/stretch bands, hydraulic machines or body weight (e.g. push-ups, chin-ups). RT has long been considered an important activity for adults and the latest physical activity recommendations for youth, adults and older adults include guidelines for RT (U.S. Department of Health & Human Services, 2008). Studies have shown that supervised RT programmes do not have adverse effects in children and adolescents (Malina, 2006) and in fact can improve muscular fitness, cardiovascular fitness, body composition, bone mineral density and blood lipid profiles (Benson, Torode, & Fiatarone Singh, 2008; Faigenbaum, 2000; Lubans, Sheaman, & Callister, 2010;

Malina, 2006). While these studies have shown promising results in the short-term, longer-term programmes that employ strategies to improve exercise adherence have not been evaluated.

Due to the importance of physical fitness for the current and future health of young people, it is important that interventions use valid and reliable measures to evaluate health-related fitness (HRF). While there are some data to support the validity of field-based measures of physical fitness (Castro-Piñero, et al., 2010), appropriate reliability data for many of the HRF tests commonly used with adolescents is lacking (Ortega, et al., 2008). There are three important types of reliability; test-retest correlation, change in mean and within subject variation (Hopkins, Hawley, & Burke, 1999). Correlation is the most commonly measured reliability and represents how well the rank order of participants in one trial is replicated in subsequent trials (Hopkins, 2000a). It has been suggested that intraclass correlation (ICC) should be cited in reliability studies, but not employed as the only statistic, since these values are highly influenced by the range of the measured values (Atkinson & Nevill, 1998). Change in the mean represents the difference between trial results and consists of systematic bias and random change. Systematic bias refers to differences between trials that may be explained by some aspect of the testing procedure. For example, participants may score higher on subsequent tests due to a learning effect or exhibit lower scores due to fatigue or boredom. Random change is a result of the random error of measurement and is larger in smaller samples (Hopkins, 2000a).

Within subject variation is considered to be the most important reliability type in sport science studies and there is debate as to whether typical error or limits of agreement (LOA) is the best measure (Atkinson & Nevill, 1998; Hopkins, 2000a; Hopkins, Marshall,

Batterham, & Hanin, 2009). Hopkins and colleagues (Hopkins, 2000a; Hopkins, et al., 2009) support the use of typical error expressed as a coefficient of variation (CV) primarily because it is not dependent on sample size and it is an easier concept to apply to research problems. Alternatively, Atkinson and Nevill (1998) endorse the use of the LOA method because it does not assume the presence or absence of heteroscedasticity in the sample data. In interpreting data from RT and HRF programmes, it is imperative that reliable measures are used before drawing conclusions about programme effectiveness. The primary aim of this study was to determine the one-week test-retest reliability of a battery of field-based HRF tests.

Interventions to engage youth in RT are clearly warranted (Stratton, et al., 2004; U.S. Department of Health & Human Services, 2008) and should be guided by a relevant theory of behaviour change (Lubans, Foster, & Biddle, 2008). Theory-based interventions are hypothesized to influence the relevant cognitions, which then mediate changes in the targeted behaviour (Hardeman, et al., 2005). Bandura's Social Cognitive Theory (Bandura, 1986) has been successfully applied to the design and evaluation of health behaviour interventions among youth (Lubans & Morgan, 2008; Lubans, Morgan, Callister, & Collins, 2009; Salmon, Ball, Hume, Booth, & Crawford, 2008). While self-efficacy and outcome expectancy have been identified as the central tenets of Bandura's theory (Bandura, 2004), valid and reliable scales for assessing RT cognitions are not currently available for youth. A secondary aim of this study was to evaluate the reliability of psychosocial scales for RT self-efficacy and outcome expectancy.

Methods

Participants

Study approval was sought and obtained from the University of Newcastle Research Ethics Committee, Newcastle and Maitland Catholic Schools Diocese and the school principal from one independent secondary school in Newcastle, New South Wales (NSW), Australia. Information letters, parental and participant consent forms were sent home with students and those who returned signed consent forms were permitted to participate in the study. Eligible participants were 68 (42 boys and 26 girls) year 9 secondary school students at the study school.

Measures

All assessments were completed by four trained research assistants and measurements were completed at the study school using the same instruments on two occasions seven days apart (Trial 1 and Trial 2, hereafter called T2 and T1). On functional tests, a 7-day separation period is commonly used (Mota, et al., 2010). A period of one week was considered sufficient time to reduce the learning effect of the testing procedures without introducing additional error due to maturation. The following tests were selected as they represent feasible field-based approaches to the assessment of body composition and muscular fitness in adolescent populations.

Stature and body mass. Body mass was measured in light clothing without shoes using a portable digital scale (Seca 770, Wedderburn) to the nearest 0.1kg and stature was measured to the nearest 0.1 cm using a portable stadiometer (Design No. 1013522, Surgical and Medical Products, Seven Hills, Australia). Body mass index (BMI) was calculated using the standard equation (body mass[kg]/stature[m]^2) and age-specific cut-off points from the International Obesity Task Force were used to classify participants as healthy weight, overweight or obese (Cole, Bellizzi, Flegal, & Dietz, 2000).

Bio-electrical impedance analysis (BIA). Percentage body fat, was determined using the Imp™ SFB7 bio-electrical impedance analyser. Participants were asked to refrain from physical activity before testing and to maintain normal hydration patterns.

Leg dynamometer. A leg dynamometer (TTM Muscular Meter, Gloria, Tokyo, Japan) was used to assess participants' lower body muscular fitness. Participants were instructed to stand upright on the base of the dynamometer with their knees bent at approximately 110 degrees and feet shoulder width apart holding onto the centre of the bar with both hands. The participants are then instructed to pull the bar as hard as possible without bending their back and keeping their arms straight.

90-degree push-up test. The 90º push-up test (90PU) was used as a measure of upper body muscular fitness (Cooper Institute for Aerobics Research, 1992). Participants started in the push up position with their hands and toes touching the floor and the arms at shoulder width apart. Keeping their back and knees straight, participants then lowered themselves to the ground until there was a 90-degree angle at the elbows, with upper arms parallel to the floor. A tester held their hand at the point of the 90-degree angle so that the shoulder of the participant being tested touched the tester's hand, then back up. The push-ups were completed in time to a metronome set at 40 beats per minute with one complete push-up every three seconds. The participant continued until they could do no more in rhythm (e.g. did not complete the last three efforts in rhythm).

7-stage abdominal strength test. The 7-stage abdominal strength test was used to provide a measure of abdominal muscular fitness. Participants started on their back, with their knees at right angles and feet flat on the floor. Starting at level 1, participants attempted to perform one complete sit-up for each of the levels (three attempts at each level are permitted). A level is achieved if a single sit up is performed in the prescribed manner, without the feet coming off

the floor. The 8 levels are as follows: Level 0- cannot perform Level 1, Level 1- with arms extended, participant curls up so that the wrists reach the knees, Level 2- with arms extended, participant curls up so that the elbows reach the knees, Level 3- with arms extended, participant curls up so that the chest touches the knees, Level 4- with arms held across the chest, holding the opposite shoulders, the participant curls up so that the forearms touch the thighs, Level 5- with the hands held behind the head, the participant curls up so that the chest touches the thighs, Level 6- uses the same technique used in Level 5 however the participant completes the curl-up with a 5lb (2.5kg) weight held behind the head, Level 7- using the same technique from Level 5 and 6 but using a 10lb (5kg) weight.

Timed wall squat test. The timed wall squat was used to assess participants' lower body muscular endurance. Participants stood on both feet with their back to a wall (hip and knees flexed to a 90° angle) and lift one foot 5cm off the ground. Timing started when the foot is lifted off the ground and ended when the foot retouches the ground. The test is repeated for each foot and scores recorded for the actual foot on the ground (i.e. the left foot is recorded while the right foot is off the ground).

Social-cognitive assessments. The secondary aim of this study was to evaluate the repeatability of two previously developed scales designed to assess adolescents' RT self-efficacy and outcome expectancies. RT self-efficacy refers to an individual's confidence to complete RT. RT outcome expectancy refers to an individual's beliefs about an individual's perceptions of the outcomes of RT. The Both scales were rated on 5-point Likert scales (1 = *Strongly disagree* to 5 = *Strongly agree*). The RT self-efficacy scale included 4-items: i) *I have the strength to complete resistance training exercises*, ii) *I can complete resistance training exercises without the help of someone else (e.g. friend, trainer),* iii) *If I don't have access to a gym I can still do resistance training (e.g. body*

weight exercises) and iv) *I have the skill and technique to complete resistance training exercises safely.* The RT outcome expectancy scale included 4-items with the common stem "*If I participate in resistance training on a regular basis (2-3 times/week) then…*". Items included: i) *It would increase my muscular strength,* ii) *It would help improve my body shape,* iii*) It would improve my performance in sports, dance, or other physical activities* and iv) *It would help put me in a better mood.* The two scales were developed through a series of phases which included: identifying potential items from the existing adult-based literature (Plotnikoff, Trinh, Courneya, Karunamuni, & Sigal, 2009); pre-pilot work with 6 adolescent males and females; and a subsequent pilot study with 106 adolescents (Lubans, Aguiar, & Callister, 2010). The scale structures in the pilot test were tested using confirmatory factor analysis in AMOS. A covariance matrix of the RT self-efficacy and outcome expectancy items found that the data were a good fit to the hypothesized two-factor model, $\chi^2 = 34.83$ ($df = 34$), $p = 0.115$, goodness of fit index = 0.92 and root mean square error of approximation = 0.07 (Bollen, 1989; Jöreskog & Sörbom, 1993). The Cronbach alphas for self-efficacy and outcome expectancy were α = 0.75 and $\alpha = 0.83$, respectively. The factor loadings were significant at $p < 0.001$ and the standardised loadings ranged from 0.59 to 0.78 for the self-efficacy scale and 0.53 to 0.83 for the outcome expectancy scale. Following the piloting process, minor changes were made to the scales for the current study.

Analysis

Statistical analyses were completed using PASW Statistics 17 (SPSS Inc. Chicago, IL) software and alpha levels were set at $p < 0.05$. Three types of reliability were assessed, including rank order repeatability, change in mean and within subject variation. Intraclass

correlation (ICC) was calculated to provide an estimate of rank order repeatability and paired samples t-tests were calculated to determine systematic and random change in mean values. Bivariate correlations between the inter-trial difference (T2-T1) and the mean of the trials [(T2-T1)/2)] were used to identify proportional bias. While LOA has been previously recommended by experts in the field (Atkinson & Nevill, 1998), more recently the use of LOA has been discouraged (Hopkins, et al., 2009). Consequently, typical error complements the other tests of reliability and was used to explore within-subject variation in the study sample (Hopkins, 2000b). However, if proportional bias was identified and could not be resolved using log transformations, the 95% LOA were calculated and reported.

Although it has been arbitrarily suggested that less than 10% variability is an acceptable level of error (Stokes, 1985), researchers should consider the purpose of their measures before determining the appropriate degree of error (Atkinson & Nevill, 1998). Faigenbaum (2000) has previously noted that strength gains of 30-40% are typically observed in untrained children following short-term RT programmes. As strength gains made by adolescent as a result of RT programmes are often smaller than this (Lubans, Sheaman, et al., 2010; Shaibi, et al., 2006; Velez, Golem, & Arent, 2010), it was decided that less than 20% variability was an acceptable degree of error. A change in body fat of 2% was considered clinically significant and achievable in short-term RT studies (Lubans, Sheaman, et al., 2010; Velez, et al., 2010) and therefore 20% variability was identified as the cut-point for % body fat. Determining what is acceptable variability for self-report variables is more challenging and previous studies have found it difficult to detect changes in psychosocial constructs (Lubans, et al., 2008). Less than 10% error was

considered to be appropriate for self-report variables to maximize the chances of detecting significant changes in these constructs.

Results

Participants (mean age = 14.8 ± 0.4 years) were 42 boys and 26 girls who completed the study assessments on two occasions separated by one week. Mean values, standard deviations and inter-trial differences for all of the variables are reported in Table 1. Based on the IOTF cut-points 33 (78.6%) and 22 (84.6%) of the boys and girls were classified as healthy weight. Six boys and six girls were classified as overweight and three boys were considered to be obese.

Insert Table 1 here

The ICCs, change in mean, bivariate correlations [between the difference (T2-T1) and the mean (T2-T1)/2)], and typical error are reported in Table 2 for boys and Table 3 for girls. There was significant heteroscedasticity for boys' leg dynamometer results (r = 0.32, p = 0.04). However, after log transforming the data, the relationship was no longer significant (r = 0.22, p = 0.18). There was also significant proportional bias for the girls' leg dynamometer (r = -0.52, p = 0.01) and wall squat (r = 0.45, p = 0.05) results. However, log transformations did not remove the heteroscedasticity.

The BIA body fat % ICC values were similar for boys (ICC = 0.95) and girls (ICC = 0.93), but the coefficient of variation (CV) was considerably higher for boys (22.2% versus 12.2%). The boys' CV values for the tests of muscular fitness ranged from 9% for the leg dynamometer test to 26.5% for the timed wall squat test. Girls' CV values were 21.4% and 17.1% for the push-up and sit-up tests, respectively. Due to data heteroscedasticity, typical errors for the leg dynamometer and wall squat tests could not

be calculated. LOA for the leg dynamometer and the wall squat test were -15.6 to 26.7 and -31.8 to 38.4, respectively.

Insert Tables 2 and 3 here

The ICC values for boys RT self-efficacy and outcome expectations were both 0.81. Girls ICC values for RT self-efficacy and outcome expectations were 0.88 and 0.69, respectively. For both boys and girls the typical errors and CVs were small and similar for RT self-efficacy and outcome expectations. There were no significant changes in mean values for boys or girls.

Discussion

The primary aim of this study was to determine the test-retest reliability of a battery of field-based tests for the evaluation of HRF and RT programmes in adolescent populations. The Imp™ SFB7 BIA machine produced acceptable error estimates in the study population but the coefficient of variation was higher among adolescent boys. The tests of muscular fitness were more reliable among girls and evidence of systematic bias was found in three of the tests of muscular fitness (i.e. leg dynamometer, push-up and wall squat tests) in boys. The secondary aim of this study was to evaluate the reliability of psychosocial scales for RT self-efficacy and outcome expectancy. Both scales demonstrated appropriate variability in the study sample, indicating that they could be included in studies evaluating the effects of RT programmes on cognitions in youth and in cross-sectional studies. Due to the disparity between the ICC and typical error results, this study has highlighted the importance of using appropriate tests of reliability for HRF and psychosocial outcomes.

The decision as to whether or not a measure is appropriate for use in a particular population is a scientific judgment and cannot be based on statistics alone (Atkinson &

Nevill, 1998; Bland & Altman, 1986; Hopkins, et al., 2009; Ortega, et al., 2008). Although researchers previously supported the use of LOA as a measure of reliability for the sport and exercise sciences (Atkinson & Nevill, 1998), more recently, typical error has been recommended by experts in the field (Hopkins, et al., 2009). Correlation should be reported in reliability studies, but not as the sole measure as it does not assess systematic bias and is largely dependent on the range of values in the sample (Atkinson & Nevill, 1998; Bates, Zhang, Dufek, & Chen, 1996). The limitations of using correlation as a measure of reliability were highlighted in the current study with lower ICC values found among girls, due to the smaller number of females in the study sample. Typical error results provide meaningful information for researchers and in the current study the measures were considered to have acceptable reliability if they were capable of detecting short-term improvements in HRF resulting from RT programmes in adolescents. In the current study, some of the measures (e.g. BIA machine and push-up test) resulted in high ICC (> 0.90) and high CV (15%) values. A wide range of scores is necessary to achieve high ICC results, but the high CV results suggest that individuals were not consistent in their test-retest results from Trial 1 to Trial 2.

The tests of muscular fitness were more reliable among girls and the sit-up test was the only test that did not significantly increase from Trial 1 to Trial 2 among boys. There was evidence of systematic bias for the leg dynamometer, push-up and wall squat tests among boys and for the push-up test for girls. The improvements from Trial 1 to Trial 2 may be explained by a learning effect and future studies should include a more extensive familiarization process to reduce the error associated with these tests. There was some evidence of proportional bias which was resolved for boys, but could not be resolved for girls. Consequently, LOA were used to test within subject variation for girls'

leg dynamometer and wall squat test results. In a recent study, Ortega et al (2008) evaluated the repeatability of the standing broad jump and the squat jump in adolescents. Both tests were found to have excellent reliability and the inter-trial differences were close to zero. However, due to the explosive nature of the movements involved in both tests, they are more a measure of power than lower body muscular strength. While the 1 RM back squat or leg press may be considered the criterion measures of lower body muscular strength and can be completed safely in the presence of qualified instructors/supervisors (Faigenbaum, Milliken, & Westcott, 2003), these tests require access to weight training facilities.

The ICC results for the push-up test from the current study are comparable to previous studies examining the rank order repeatability of this test in youth (McManis, Baumgartner, & Wuest, 2000; Pate, Burgess, Woods, Ross, & Baumgartner, 1993). For example, McManis et al (2000) tested the reliability of the push-up test in adolescents and found ICC values ranged from 0.50 to 0.86. However, the authors did not report typical error or LOA and there was no consistency in the time period between trials as the second tests were conducted three to seven days from the first day of testing. Pate and colleagues (Pate, et al., 1993) found the Chrysler Fund-Amateur Athletic Union push-up test to have a test-retest reliability coefficient of 0.85. The same authors also examined the relationship between various tests of upper body muscular strength and a one repetition maximum (RM) bench press (criterion measure of strength) and found the push-up test to have the strongest correlation to the criterion measure. In the current study, the typical error of the push-up test was lower in girls than boys, but the CV was higher in girls. For both groups there was evidence of systematic error, suggesting that a more extensive preparation period is required before testing.

The CV for the 7-stage abdominal sit-up test was 12.5% and 17.1% for boys and girls, respectively. Anderson and colleagues (1997) tested the reliability of the FITNESSGRAM abdominal strength test and a timed curl-up test over a two day period. The ICCs for both their measures were above 0.70, however, the authors did not report typical error or LOA. While the ICC for the wall squat test in our study ranged from 0.69 to 0.88, the CV was unacceptably high for both boys and could not be calculated for girls. This finding reinforces the importance of reporting typical error or LOA for fitness measures. The variability associated with this test makes it inappropriate for use with adolescents.

The typical error of the Imp™ SFB7 BIA machine for boys and girls was considered acceptable. Although the CV was considerably higher for boys, this finding may be explained by the homogeneity of the boys in the study sample, who were relatively lean (mean body fat = 9.7 ± 6.7%). Because the mean % body fat of girls in the study sample was twice the boys' mean, the girls' CV for this measures was much lower. Alternative explanations for the variability of the BIA machine results include inconsistent hydration across trials and lack of precision in electrode placement, both of which are important considerations when assessing body composition using bio-electrical impedance (Nielsen, et al., 2007). Studies have shown that RT is an effective strategy for improving body composition in adolescents (Lubans, Morgan, Callister, Collins, & Plotnikoff, 2010; Shaibi, et al., 2006; Velez, et al., 2010) and BIA is an attractive method for assessing body composition in this group. BIA is a safe, easy to use, non-invasive technique which has been found to accurately predict whole body fat free mass in youth (Jensky-Squires, et al., 2008; Nielsen, et al., 2007). For example, Jensky-Squires and colleagues (2008) found that body fat % as calculated by the BioSpace InBody 320 was

reliable and significantly correlated with under water weighing in youth. As there are many BIA devices available, which use different equations to calculate body fat, it is important that researchers continue to publish the validity and reliability of these instruments as they are made available.

RT self-efficacy and outcome expectancy were identified for inclusion in the study because they represent important cognitions that are positively associated with both RT and aerobic physical activity behaviour (Biddle & Fuchs, 2009; Lubans, et al., 2008; Plotnikoff, et al., 2009; Van der Horst, Paw, Twisk, & Van Mechelen, 2007). Self-efficacy refers to situation specific physical activity and according to Bandura (1986) there are four main sources which influence an individual's self-efficacy. These include prior success and performance attainment, imitation and modelling, verbal and social persuasion, and judgments on physiological states and RT programmes should explicitly target these sources to improve adolescents' confidence. Outcome expectancy has played a crucial role in the development of cognitive explanations of behaviour (Williams, Anderson, & Winett, 2005) and considering the myths and misconceptions surrounding RT (Faigenbaum, et al., 2009), educating adolescents about the benefits of RT may be an essential programme component necessary to increase exercise adherence among youth.

The RT self-efficacy and outcome expectancy scales both demonstrated adequate test-retest reliability in the study sample. Few studies have reported the psychometric properties of RT-specific psychosocial scales in adolescents and the scales tested in the current study might be used to evaluate the impact of RT programmes on psychosocial outcomes in future studies. Interventions designed to

encourage RT are warranted and studies should be guided by a relevant theory of behaviour change (Michie & Abraham, 2004). The assessment of relevant and reliable cognitions will help improve our understanding of behaviour change and exercise adherence in interventions.

There are some study limitations that should be noted. A larger sample size and the inclusion of an additional trial may have helped to improve the precision of the reliability estimates. We were not able to examine the results by age group and our small and relatively homogenous sample may have contributed to our high CV results. Finally, we identified systematic error in a number of the tests and future studies should include more extensive practice periods to reduce this error.

To the authors' knowledge no previous study has published the reliability of the Imp™ SFB7 tetra polar bio-electrical impedance spectroscopy device in an adolescent population. While previous studies have reported the test-retest correlation of the leg dynamometer, push-up, sit-up and wall squat tests, no previous study has reported the typical error or CV for these tests in an adolescent population. The high levels of systematic error suggest that these measures may require an extensive familiarisation phase before the results can be considered reliable.

References

Anderson, E. A., Zhang, J. J., Rudisill, M. E., & Gaa, J. (1997). Validity and reliability of a timed curl-up test: Development of a parallel form for the FITNESSGRAM abdominal strength test. *Research Quarterly for Exercise & Sport, 68*(1), A51.

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for addressing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine, 26,* 217-238.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, N.J: Prentice-Hall.

Bandura, A. (2004). Health promotion by social cognitive means. *Health Education & Behavior, 31,* 143-164.

Bates, B. T., Zhang, S., Dufek, J. S., & Chen, F. C. (1996). The effects of sample size and variability on the correlation coeffficient. *Medicine & Science in Sport & Exercise, 28*(3), 386-391.

Benson, A. C., Torode, M. E., & Fiatarone Singh, M. A. (2008). Effects of resistance training on metabolic fitness in children and adolescents: a systematic review. *Obesity Reviews, 9,* 43-66.

Biddle, S. J. H., & Fuchs, R. (2009). Exercise psychology: A view from Europe. *Psychology of Sport & Exercise, 10,* 410-419.

Blair, S. N., Cheng, Y., & Holder, J. S. (2001). Is physical activity or physical fitness more important in defining health benefits? *Medicine & Science in Sports & Exercise, 33*(6), S379-S399.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* 307-310.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Boreham, C. A., & Riddoch, C. (2001). The physical activity, fitness and health of children. *Journal of Sport Sciences, 19*(12), 915-929.

Castro-Piñero, J., Artero, E. G., España-Romero, V., Sjöström, M., Suni, J., & Ruiz, J. R. (2010). Criterion-related validity of field-based fitness tests in youth: a systematic review. *British Journal of Sports Medicine, 44*, 934-943.

Cole, T. J., Bellizzi, M. C., Flegal, K. M., & Dietz, W. H. (2000). Establishing a standard definition for child overweight and obesity worldwide: international survey. *British Medical Journal, 320*(7244), 1240-.

Cooper Institute for Aerobics Research (1992). *The Prudential FITNESSGRAM: Test administration*. Dallas, TX: Cooper Institute for Aerobics Research.

Ekelund, U., Poortvliet, E., Nilsson, A., Yngve, A., Holmberg, A., & Sjostrom, M. (2001). Physical activity in relation to aerobic fitness in 14- to 15-year-old boys and girls. *European Journal of Applied Physiology, 85*(195-201).

Faigenbaum, A. D. (2000). Strength training for children and adolescents. *Clinical Sports Medicine, 19*(4), 593-619.

Faigenbaum, A. D., Kraemer, W. J., Blimkie, C. J., Jeffreys, I., Micheli, L. J., Nitka, M., et al. (2009). Youth resistance training: updated position statement paper from the national strength and conditioning association. *Journal of Strength & Conditioning Research, 23*(5 Suppl), S60-79.

Faigenbaum, A. D., Milliken, L. A., & Westcott, W. L. (2003). Maximal strength testing in healthy children. *Journal of Strength & Conditioning Research, 17*(1), 162-166.

Hardeman, W., Sutton, S., Griffin, S., Johnston, M., White, A., Wareham, N., et al. (2005). A causal modelling approach to the development of theory based behavior change programmes for trial evaluation. *Health Education Research,, 20*, 676-687.

Hopkins, W. G. (2000a). Measures of reliability in sports medicine and science. *Sports Medicine, 30*(1), 1-15.

Hopkins, W. G. (2000b). Reliability from consecutive pairs of trials (Excel spreadsheet). In: A new view of statistics. sportsci.org: Internet Society for Sport Science, sportsci.org/resource/stats/xrely.xls.

Hopkins, W. G., Hawley, J. A., & Burke, L. M. (1999). Design and analysis of research on sport performance enhancement. *Medicine & Science in Sports & Exercise, 31*, 472-485.

Hopkins, W. G., Marshall, S. W., Batterham, A. M., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine & Science in Sport & Exercise, 41*(1), 3-12.

Jensky-Squires, N. E., Dieli-Conwright, C. M., Rossuello, A., Erceg, D. N., McCauley, S., & Schroeder, E. T. (2008). Validity and reliability of body composition analysers in children and adults. *British Journal of Nutrition, 100*(4), 859-865.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: user's reference guide*. Chicago: Scientific Software International.

Lubans, D. R., Aguiar, E., & Callister, R. (2010). The effects of free weights and elastic tubing resistance training on physical self-perception in adolescents. *Psychology of Sport & Exercise, 11*(6), 497-504.

Lubans, D. R., Foster, C., & Biddle, S. J. H. (2008). A review of mediators of behavior in interventions to promote physical activity among children and adolescents. *Preventive Medicine, 47,* 463-470.

Lubans, D. R., & Morgan, P. J. (2008). Evaluation of an extra-curricular school sport program promoting lifestyle and lifetime activity. *Journal of Sport Sciences, 26*(5), 519-529.

Lubans, D. R., Morgan, P. J., Callister, R., & Collins, C. E. (2009). Effects of integrating pedometers, parental materials, and email support within an extracurricular school sport intervention. *Journal of Adolescent Health, 44*(2), 176-183.

Lubans, D. R., Morgan, P. J., Callister, R., Collins, C. E., & Plotnikoff, R. A. (2010). Exploring the mechanisms of physical activity and dietary behavior change in the Program X intervention for adolescents. *Journal of Adolescent Health, 47*(1), 83-91.

Lubans, D. R., Sheaman, C., & Callister, R. (2010). Exercise adherence and intervention effects of two school-based resistance training programs for adolescents. *Preventive Medicine, 50*(1), 56–62.

Malina, R. M. (2006). Weight training in youth-growth, maturation, and safety: an evidence-based review. *Clinical Journal of Sports Medicine, 16*(6), 478-487.

McManis, B. G., Baumgartner, T. A., & Wuest, D. A. (2000). Objectivity and reliability of the 90 degree push-up test. *Measurement in Physical Education & Exercise Science, 4*(1), 57-67.

Michie, S., & Abraham, C. (2004). Interventions to change health behaviours: Evidence-based or evidence-inspired? *Psychology & Health, 19*(1), 29-49.

Mota, J., Vale, S., Martins, C., Gaya, A., Moreira, C., Santos, R., et al. (2010). Influence of muscle fitness test performance on metabolic risk factors among adolescent girls. *Diabetology & metabolic syndrome, 23*(2), 42.

Nielsen, B. M., Dencker, M., Ward, L., Linden, C., Thorsson, O., Karlsson, M. K., et al. (2007). Prediction of fat-free body mass from bioelectrical impedance among 9- to 11-year-old Swedish children. *Diabetes, obesity & metabolism, 9*(9), 521-539.

Ortega, F. B., Artero, E. G., Ruiz, J. R., Vicente-Rodriguez, G., Bergman, P., Hagstro¨mer, M., et al. (2008). Reliability of health-related physical fitness tests in European adolescents. The HELANA study. *International Journal of Obesity, 32*, S49-S57.

Pate, R. R., Burgess, M. L., Woods, J. A., Ross, J. G., & Baumgartner, T. (1993). Validity of field tests of upper body muscular strength. *Research Quarterly for Exercise & Sport, 64*(1), 17-24.

Plotnikoff, R., Trinh, L., Courneya, K., Karunamuni, N., & Sigal, R. (2009). Predictors of aerobic physical activity and resistance training among Canadian adults with type 2 diabetes: An application of the Protection Motivation Theory. *Psychology of Sport & Exercise, 10*(3), 320-328.

Rizzo, N. S., Ruiz, J. R., Hurtig-Wennlof, A., Ortega, F. B., & Sjostrom, M. (2007). Relationship of physical activity, fitness, and fatness with clustered metabolic risk in children and adolescents: the European Youth Heart Study. *Journal of Pediatrics, 150*(4), 388-394.

Ruiz, J. R., Castro-Piñero, J., Artero, E. G., Ortega, F. B., Sjöström, M., Suni, J., et al. (2009). Predictive validity of health-related fitness in youth: a systematic review. *British Journal of Sports Medicine, 43*, 909-923.

Salmon, J., Ball, K., Hume, C., Booth, M., & Crawford, D. (2008). Outcomes of a group-randomized trial to prevent excess weight gain, reduce screen behaviors and promote physical activity in 10-year-old children: Switch-Play. *International Journal of Obesity, 32,* 601-612.

Shaibi, G. Q., Cruz, M., Ball, G., Weigensberg, M. J., Salem, G. J., Crespo, N. C., et al. (2006). Effects of resistance training on insulin sensitivity in overweight Latino adolescent males. *Medicine & Science in Sports & Exercise, 38*(7), 1208-1215.

Stokes, M. (1985). Reliability and repeatability of methods for measuring muscle in physiotherapy. *Physiotherapy Practice, 7*(1), 71-76.

Stratton, G., Jones, M., Fox, K. R., Tolfrey, K., Harris, J., Maffulli, N., et al. (2004). BASES position statement on guidelines for resistance exercise in young people. *Journal of Sports Sciences, 22,* 383-390.

U.S. Department of Health & Human Services (2008). *2008 Physical Activity Guidelines for Americans*. Washington, D.C.: U.S. Department of Health & Human Services.

Van der Horst, K., Paw, M. J. C. A., Twisk, J. W. R., & Van Mechelen, W. (2007). A brief review on correlates of physical activity and sedentariness in youth. *Medicine & Science in Sport & Exercise, 39*(8), 1241-1250.

Velez, A., Golem, D. L., & Arent, S. M. (2010). The impact of a 12-week resistance training program on strength, body composition, and self-concept of Hispanic adolescents. *Journal of Strength & Conditioning Research, 24*(4), 1065-1073.

Williams, D. M., Anderson, E. S., & Winett, R. A. (2005). A review of the outcome expectancy construct in physical activity research. *Annals of Behavioral Medicine, 29*(1), 70-79.

Table 1: Results from physical fitness tests and resistance training beliefs scales (mean ± SD) in boys (n = 42) and girls (n = 26)

Variables	1st Trial (T1)		2nd Trial (T2)		Inter-trial difference (T2 –T1)		
	Boys	Girls	Boys	Girls	Boys	Girls	Total
Body composition							
BIA (body fat %)	9.7 ± 6.7	18.6 ± 5.4	10.0 ± 7.5	17.6 ± 6.5	0.3 ± (3.1)	-0.9 ± 3.1	-0.1 ± 3.1
Muscular fitness							
Leg dynamometer (kg)	104.5 ± 21.0	74.0 ± 16.6	111.0 ± 25.1	77.2 ± 10.8	6.4 ± 13.4	3.4 ± 11.9	5.2 ± 12.8
Push-up test (reps)	22.2 ± 7.5	10.5 ± 6.7	24.9 ± 8.8	12.9 ± 6.9	2.7 ± 5.0	2.3 ± 3.5	2.6 ± 4.5
Sit-up test (level)	4.8 ± 1.4)	3.9 ± 1.8	4.8 ± 1.2	4.2 ± 1.5	0.0 ± 0.9	0.3 ± 0.9	0.1 ± 0.9
Wall squat test (secs)[a]	42.5 ± 24.6	27.6 ± 14.6	50.4 ± 28.3	30.9 ± 21.8	7.8 ± 17.4	3.3 ± 17.9	6.2 ± 17.6
Self-report scales							
Self-efficacy	4.1 ± 0.5	4.2 ± 0.5	4.1 ± 0.4	4.2 ± 0.4	0.0 ± 0.4	0.0 ± 0.3	0.0 ± 0.4
Outcome expectancy	4.2 ± 0.5	4.1 ± 0.4	4.2 ± 0.5	4.2 ± 0.3	0.1 ± 0.3	0.0 ± 0.3	0.0 ± 0.3

Note. Means and standard deviations reported; ICC = intra class correlation; BIA = bioelectrical impedance; CIs = confidence intervals
[a]The average of the left and right sides is reported.

Table 2: Reliability of physical fitness tests and resistance training scales in adolescent boys

Variables	ICC (95% CIs)	Change in mean		*r* [a]	Within subject variation[b]	CV (%)
		Mean ± (SD)	*p*			
Anthropometrics						
BIA (body fat %)	0.95 (0.90 to 0.97)	0.3 ± (3.1)	0.600	0.27	2.2 (1.8 to 2.8)	22.2
Muscular fitness						
Leg dynamometer (kg)	0.91 (0.83 to 0.95)	6.4 ± 13.4	0.005**	0.32*	9.0 (7.0 to 12.4)	-
Push-up test (reps)	0.90 (0.80 to 0.95)	2.7 ± 5.0	0.001**	0.28	3.6 (2.9 to 4.6)	15.3
Sit-up test (level)	0.96 (0.74 to 0.93)	0.0 ± 0.9	1.000	-0.22	0.6 (0.5 to 0.8)	12.5
Wall squat test (s)	0.88 (0.76 to 0.94)	7.8 ± 17.4	0.012*	0.23	12.3 (9.9 to 16.1)	26.5
Self-report scales						
Self-efficacy	0.81 (0.65 to 0.90)	0.0 ± 0.4	0.323	-0.09	0.25 (0.20 to 0.31)	6.1
Outcome expectancy	0.81 (0.64 to 0.90)	0.1 ± 0.3	0.916	0.02	0.26 (0.21 to 0.33)	6.2

Note. SD = standard deviations; ICC = intra class correlation; BIA = bioelectrical impedance; CIs = confidence intervals; s = seconds; CV = coefficient of variation
[a]Bivariate correlations between the difference (T2-T1) and the mean [(T2-T1)/2)]. If significant heteroscedasticity was identified variables were log transformed.
[b]Typical error used to identify within subject variation. For log transformed data, typical error expressed as a CV (%) is reported.
*$p < 0.05$, ** $p < 0.01$.

Table 3: Reliability of physical fitness tests and resistance training scales in adolescent girls

Variables	ICC (95% CIs)	Change in mean		*r* [a]	Within subject variation[b]	CV (%)
		Mean ± (SD)	*p*			
Anthropometrics						
BIA (body fat %)	0.93 (0.83 to 0.97)	-0.9 ± 3.1	0.179	0.38	2.2 (1.7 to 3.1)	12.2
Muscular fitness						
Leg dynamometer (kg)	0.78 (0.49 to 0.90)	3.4 ± 11.9	0.179	-0.52**	-15.6 to 26.7[c]	-
Push-up test (reps)	0.93 (0.84 to 0.97)	2.3 ± 3.5	0.002**	0.05	2.5 (2.0 to 3.5)	21.4
Sit-up test (level)	0.91 (0.81 to 0.96)	0.3 ± 0.9	0.148	0.33	0.7 (0.5 to 0.9)	17.1
Wall squat test (s)	0.69 (0.23 to 0.88)	3.3 ± 17.9	0.425	0.45*	-31.8 to 38.4[c]	-
Self-report scales						
Self-efficacy	0.88 (0.73 to 0.95)	0.0 ± 0.3	0.746	-0.22	0.2 (0.2 to 0.3)	4.8
Outcome expectancy	0.69 (0.29 to 0.86)	0.0 ± 0.3	0.847	-0.12	0.2 (0.2 to 0.3)	4.9

Note. SD = standard deviations; ICC = intra class correlation; BIA = bioelectrical impedance; CIs = confidence intervals; s = seconds; CV = coefficient of variation
[a]Bivariate correlations between the difference (T2-T1) and the mean [(T2-T1)/2)]. If significant heteroscedasticity was identified variables were log transformed.
[b]Typical error used to identify within subject variation.
[c]Log transformations did not remove heteroscedasticity and 95% limits of agreement are reported (inter-trial mean difference ± 1.96 SDs)
*$p < 0.05$, ** $p < 0.01$.