International Journal of Obesity (2008) 32, S49–S57

0307-0565/08 $32.00
www.nature.com/ijo

ORIGINAL ARTICLE

Reliability of health-related physical fitness tests in European adolescents. The HELENA Study

FB Ortega[1,2], EG Artero[1], JR Ruiz[2], G Vicente-Rodriguez[3], P Bergman[2], M Hagströmer[2], C Ottevaere[4], E Nagy[5], O Konsta[6], JP Rey-López[3], A Polito[7], S Dietrich[8], M Plada[9], L Béghin[10], Y Manios[6], M Sjöström[2] and MJ Castillo[1], on behalf of the HELENA Study Group[11]

[1]*Department of Physiology, School of Medicine, University of Granada, Granada, Spain;* [2]*Unit for Preventive Nutrition, Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden;* [3]*Escuela Universitaria de Ciencias de la Salud, Universidad de Zaragoza, Domingo Miral, Zaragoza, Spain;* [4]*Unit of Nutrition and Food Safety, Faculty of Medicine and Health Sciences, Department of Public Health, Ghent University, Ghent, Belgium;* [5]*Medical Faculty, Department of Paediatrics, University of Pécs, Pécs, Hungary;* [6]*Department of Nutrition and Dietetics, Harokopio University, Athens, Greece;* [7]*INRAN, National Research Institute for Food and Nutrition, Roma, Italy;* [8]*Division of Clinical Nutrition, Department of Pediatrics, Medical University of Vienna, Wien, Austria;* [9]*Department of Social Medicine, School of Medicine, Preventive Medicine and Nutrition Clinic, Heraklion, Crete, Greece;* [10]*Faculté de médecine, Université de Lille 2 Droit et Santé, and CIC-9301-CHRU-INSERM de Lille IFR 114, Institut de Médicine Prédictive et Thérapeutique de Lille, Lille, France*

Objective: To examine the reliability of a set of health-related physical fitness tests used in the European Union-funded Healthy Lifestyle in Europe by Nutrition in Adolescence (HELENA) Study on lifestyle and nutrition among adolescents.
Design: A set of physical fitness tests was performed twice in a study sample, 2 weeks apart, by the same researchers.
Participants: A total of 123 adolescents (69 males and 54 females, aged 13.6 ± 0.8 years) from 10 European cities participated in the study.
Measurements: Flexibility, muscular fitness, speed/agility and aerobic capacity were tested using the back-saver sit and reach, handgrip, standing broad jump, Bosco jumps (squat jump, counter movement jump and Abalakov jump), bent arm hang, 4 × 10 m shuttle run, and 20-m shuttle run tests.
Results: The ANOVA analysis showed that neither systematic bias nor sex differences were found for any of the studied tests, except for the back-saver sit and reach test, in which a borderline significant sex difference was observed ($P = 0.044$). The Bland–Altman plots graphically showed the reliability patterns, in terms of systematic errors (bias) and random error (95% limits of agreement), of the physical fitness tests studied. The observed systematic error for all the fitness assessment tests was nearly 0.
Conclusions: Neither a learning nor a fatigue effect was found for any of the physical fitness tests when repeated. The results also suggest that reliability did not differ between male and female adolescents. Collectively, it can be stated that the reliability of the set of physical fitness tests examined in this study is acceptable. The data provided contribute to a better understanding of physical fitness assessment in young people.
International Journal of Obesity (2008) **32,** S49–S57; doi:10.1038/ijo.2008.183

Keywords: fitness; reliability; Bland–Altman; adolescents

Introduction

Health-related physical fitness includes the characteristics of functional capacity and is affected by the physical activity level and other lifestyle factors. Maintaining an appropriate level of health-related physical fitness allows a person to meet emergencies, reduce the risk of disease and injury, work efficiently, participate and enjoy physical activity (sports, recreation, leisure) and look one's physical best. A high health-related physical fitness level focuses on optimum health and prevents the onset of disease and problems associated with inactivity at all ages.[1–4]

Correspondence: Dr FB Ortega, Department of Medical Physiology, School of Medicine, University of Granada, Granada 18071, Spain.
E-mail: ortegaf@ugr.es
[11]See Appendix at the end of the supplement on page S82.

The HELENA (Healthy Lifestyle in Europe by Nutrition in Adolescence) Study[5] includes a thorough assessment of health-related physical fitness. For this purpose, a set of standardized tests has been chosen, and the scientific rationale for their selection has been published elsewhere.[6]

Reliability can be defined as the consistency of measurements. Terms that have been used interchangeably with reliability in the literature are 'repeatability', 'reproducibility', 'consistency', 'agreement', 'concordance' and 'stability'. Another related but different concept is validity. Validity is the ability of the measurement tool to measure what it is designed to measure. The validity of a tool is judged by comparison with a 'gold standard' method. Definitions and detailed discussions about reliability issues in sport sciences-related research and general science can be found in the reviews published by Atkinson and Nevill,[7] Rothwell[8] and Bruton *et al*.[9]

Realistically, some amount of error is always present when collecting data. The main components of measurement error are systematic bias (for example, general learning on the tests) and random error due to biological or mechanical variation. Several statistical methods have been used to evaluate certain aspects of reliability. Correlation analysis has been commonly used, but this has limitations that will be discussed further in this paper. The study on the agreement between two measurements by means of the Bland–Altman approach seems a more proper and useful method for reliability analyses.[7–9]

In this paper, we report the outcome of reliability testing, on a test-retest basis, of the set of health-related physical fitness tests used in the HELENA Study.[6] The outcome is discussed and compared with the outcome of an extensive overview of published data on reliability testing.

Methods

Study design

The HELENA Study (http://www.helenastudy.com) is a European Union-funded project on lifestyle and nutrition among adolescents from 10 European cities: Athens and Heraklion (Greece), Dortmund (Germany), Ghent (Belgium), Lille (France), Pécs (Hungary), Rome (Italy), Stockholm (Sweden), Vienna (Austria) and Zaragoza (Spain). The study was approved by the Research Ethics Committees of each city involved. Written informed consent was obtained from the parents of the adolescents and the adolescents themselves.

From a total sample of 204 adolescents who participated in the HELENA pilot study and performed all the physical fitness tests, a subsample of 123 adolescents (69 males and 54 females, aged 13.6 ± 0.8 years) were asked to undergo the tests again 2 weeks later. The same inter-trial period has been used earlier in similar reliability studies carried out on healthy young people.[10] The two physical fitness measurements were performed at the same time of day by the same researchers. Those adolescents who took part in the retest study did not differ in age, height, weight or body mass index (BMI) ($P > 0.05$) from those adolescents who did not do so.

Anthropometric measurements

Anthropometric measurements were made with the participants barefoot and in their underwear. Weight was measured using an electronic scale (Type SECA 861) and recorded to the nearest 100 g. Height was measured using a telescopic height measuring instrument (Type SECA 225). The instrument was calibrated before the measurements with a metal calibrating rod. Height was recorded to the last complete 1 mm.

Physical fitness assessment

An extended and detailed manual of operations was designed for and thoroughly read by every researcher involved in fieldwork before the data collection started. In addition, a workshop training week was carried out in Zaragoza (Spain) in January 2006, to standardize and harmonize the measurement of physical fitness. The field workers were asked to always perform the same fitness tests so that they would become specialized in a single fitness measurement, and to minimize the potential inter-rater variability within each centre. The instructions given to the participants in every test were standardized for all the cities and were translated into the local language. In this way, the same verbal information was given to all participants in the HELENA Study.

The health-related physical fitness components, that is, flexibility, muscular strength, speed/agility and aerobic capacity (hereafter called cardiorespiratory fitness), were assessed by the physical fitness tests described below. The scientific rationale for the selection of all of these tests has been published earlier.[6]

(1) *Back-saver sit and reach test* (flexibility assessment): a standard box with a small bar, which has to be pushed by the participant, was used to perform the test. The adolescent bends his/her trunk and reaches forward as far as possible from a seated position, with one leg straight and the other bent at the knee. The test is performed once again with the opposite leg. The farthest position of the bar reached by each leg was scored in centimetres and the average of the distances reached by both legs was used in the analysis.
(2) *Handgrip test* (maximum handgrip strength assessment): a hand dynamometer with adjustable grip was used (TKK 5101 Grip D; Takey, Tokyo, Japan). The participant squeezes gradually and continuously for at least 2 s, performing the test with the right and left hands in turn, using the optimal grip span. The handgrip span was adjusted according to hand size using the equation that we have developed specifically for adolescents.[11] The

maximum score in kilograms for each hand was recorded. The average of the scores achieved in both handgrip tests was used in the analysis.

(3) *Standing broad jump test* (lower limb explosive strength assessment): from a starting position immediately behind a line, standing with feet approximately shoulder's width apart, the adolescent jumps as far as possible with feet together. The result was recorded in centimetres. A non-slip hard surface, chalk and a tape measure were used to perform the test.

(4) *The Bosco protocol* is composed of three different jumps: (4.1) *Squat jump* (lower limb explosive strength assessment): the adolescent performs a vertical jump without rebound movements starting from a half-squat position, keeping both knees bent at 90°, the trunk straight and both hands on hips. Previous counter movements are not allowed. (4.2) *Counter movement jump* (lower limb explosive strength and elastic component assessment): in a standing position, with legs straight and both hands on hips, the adolescent performs a vertical jump with an earlier fast counter movement. (4.3) *Abalakov jump* (lower limb explosive strength, elastic component and intermuscular coordination capacity assessment): the Abalakov jump is similar to the counter movement jump, but now the adolescent is allowed to freely coordinate the arms and trunk movements to reach the maximum height. The jump height is recorded in centimetres. The Infrared Platform ERGO JUMP Plus—BOSCO SYSTEM (Byomedic, SCP, Barcelona, Spain) was used for the jump assessment.

(5) *Bent arm hang test* (upper limb endurance strength assessment): the adolescent hangs from a bar for as long as possible, with the arms bent at 90 degrees. The palms face forward and the chin must be over the bar's plane. The time spent in this position, to the nearest tenth of a second, is recorded. A cylindrical horizontal bar and a stopwatch were used to perform the test.

(6) *4 × 10 m shuttle run test* (speed of movement, agility and coordination assessment): two parallel lines are drawn on the floor 10 m apart. The adolescent runs as fast as possible from the starting line to the other line and returns to the starting line, crossing each line with both feet every time. This is performed twice, covering a distance of 40 m (4 × 10 m). Every time the adolescent crosses any of the lines, he/she should pick up (the first time) or exchange (second and third time) a sponge that has earlier been placed behind the lines. The stopwatch is stopped when the adolescent crosses the end line with one foot. The time taken to complete the test is recorded to the nearest tenth of a second. A slip-proof floor, four cones, a stopwatch and three sponges were used to perform the test.

(7) *20-m shuttle run test* (cardiorespiratory fitness assessment): the adolescents perform the test as described earlier by Léger *et al*.[12] Participants are required to run between two lines 20 m apart, while keeping pace with audio signals emitted from a pre-recorded CD. The initial speed is 8.5 km h^{-1}, which is increased by 0.5 km h^{-1} min^{-1} (1 min equals one stage). Participants are instructed to run in a straight line, to pivot on completing a shuttle, and to pace themselves in accordance with the audio signals. The test is finished when the participant fails to reach the end lines concurrent with the audio signals on two consecutive occasions. Otherwise, the test ends when the participant stops because of fatigue. All measurements were carried out under standardized conditions on an indoor rubber-floored gymnasium. The participants were encouraged to keep running as long as possible throughout the course of the test. The last completed stage or half-stage at which the participant drops out was scored. A gymnasium or space large enough to mark out a 20 m track, a 20 m tape measure, a CD player and a CD with the audio signals recorded were used to perform the test.

All the tests were performed twice and the best score was retained, except for the bent arm hang and the 20-m shuttle run tests, which were performed only once.

Review of fitness reliability studies

The search strategy for identifying the fitness reliability studies was based on combinations of the following terms: fitness, reliability, repeatability, reproducibility and measurement error. The databases used were Medline, PubMed and SportDiscus. The electronic search identified 112 publications that concerned the reliability of fitness assessment. The inclusion criteria were studies involving healthy children and/or adolescents aged 18 years or younger, and those published since 1990. In the end, 22 studies met the inclusion criteria and were selected. An additional search was carried out to find fitness reliability studies that used the Bland–Altman approach, including healthy or unhealthy people at any age.

Statistical analysis

The data are presented as means ± s.d., unless otherwise stated. Both the potential systematic bias (H_0; mean inter-trial difference = 0; H_1; mean inter-trial difference ≠ 0) and sex differences on the studied physical fitness tests were analysed by one-way analysis of variance (ANOVA) on inter-trial difference (test 2−test 1, hereafter called T2−T1) with sex as a fixed factor. As no sex-specific effect on reliability of the studied physical fitness tests was found, the analyses were performed for both males and females together. The agreement between the corresponding fitness variables obtained during the two successive measurements was also examined graphically by plotting the difference between each pair of measurements against their mean, according to the Bland and Altman approach.[13,14] The 95% limits of agreement for all the physical fitness variables were

calculated as the inter-trial mean difference ± 1.96 s.d. (of the inter-trial differences).

As the standard deviation for a sample of two observations can be written as $|T2-T1|/\sqrt{2}$, the presence of heteroscedasticity can then be analysed in line with the Bland–Altman approach by using the Kruskal–Wallis test, a non-parametric one-way ANOVA. A significant *P*-value would confirm heteroscedasticity, which means that the inter-trial variability, |T2−T1|, of a physical fitness test would differ with the physical fitness level groups. Sex-specific quartiles were estimated for every test performed and were used to classify the adolescents into different fitness levels. Distribution of the residuals for the inter-trial difference variables (T2−T1), but not for the absolute difference variables |T2−T1|, showed a satisfactory pattern. Therefore, parametric (ANOVA) and non-parametric (Kruskal–Wallis) approaches were used in this paper.

All calculations were performed using SPSS v.15.0 software for Windows. For all analyses, the significance level was 5%.

Results

The physical characteristics of the study sample are shown in Table 1. Mean values and standard deviation for the two trials, as well as the mean inter-trial difference for the physical fitness tests in the studied male and female adolescents, are also shown in Table 1. Neither systematic bias nor sex differences were found for any of the studied tests, except for the back-saver sit and reach test, in which a borderline significant sex difference was observed ($P=0.044$).

The Bland–Altman plots (Figures 1–3) graphically showed the reliability patterns, in terms of systematic errors (bias or mean inter-trial differences) and random error (95% limits of agreement), of the physical fitness tests studied. It can be observed that the systematic error when fitness assessment was performed twice was nearly 0 for all the tests.

The heteroscedasticity analysis showed that the higher the bent arm hang score (quartiles), the higher was the inter-trial difference ($P<0.001$). Moreover, it was observed that adolescents who scored high in the back-saver sit and reach test had a better inter-trial agreement compared with those adolescents who scored lower ($P<0.01$).

Discussion

Review of fitness reliability studies and methodological discussion

Table 2 summarizes the fitness reliability studies carried out in healthy young people since 1990. The most frequently used statistical approaches to assess overall agreement between measurements were correlation methods (used in 95% of the reviewed studies). However, correlation is a measure of the strength of association between two variables but not necessarily a measure of agreement. Its use is considered inappropriate for that purpose because, first, it is not possible to assess systematic bias, and second, it depends on the range of the values in the sample.[7,14] For example, if an observer always overestimates (a positive systematic bias) the 4 × 10 m shuttle run test score by 20% compared with another observer, the correlation between the measurements would be perfect, but they would never agree. Moreover, the more heterogeneous the study sample, the greater the correlation. The intraclass correlation coefficient is an appropriate overall summary measure of agreement between measurements, which reflects both systematic bias and random error in test scores.[8] However, it does not give any information on any variation in agreement with the size of the measurement, and it is also affected by the sample range.[7]

Table 1 Reliability of physical fitness tests (mean ± s.d.) in male ($n=69$) and female ($n=54$) adolescents

	1st Trial (T1)		*2nd Trial (T2)*		*Inter-trial difference (T2−T1)*	
	Males	*Females*	*Males*	*Females*	*Males*	*Females*
Age (years)	13.7 ± 0.8	13.6 ± 0.8	—	—		
Weight (kg)	56.4 ± 12.5	53.0 ± 8.9	—	—		
Height (cm)	164 ± 9.7	161 ± 6.2	—	—		
Body mass index (kg/m^2)	21.0 ± 3.8	20.5 ± 2.9	—	—		
Back-saver sit and reach (cm)[a]	19.1 ± 7.2	24.8 ± 9.1	18.8 ± 7.4	26.2 ± 8.4	−0.3 ± 3.6	1.4 ± 4.9*
Handgrip (kg)[a]	31.2 ± 6.4	26.1 ± 5.1	31.5 ± 6.9	26.1 ± 4.9	0.3 ± 2.5	0.0 ± 1.8
Standing broad jump (cm)	172 ± 29.3	147 ± 23.3	172 ± 27.0	147 ± 25.2	−0.3 ± 12.9	0.3 ± 9.0
Squat jump (cm)	24.8 ± 5.1	22.7 ± 5.0	24.3 ± 5.2	21.8 ± 3.9	−0.5 ± 3.3	−0.8 ± 3.0
Counter movement jump (cm)	26.2 ± 5.2	23.8 ± 5.0	26.1 ± 4.8	23.3 ± 4.5	0.0 ± 3.4	−0.4 ± 3.3
Abalakov jump (cm)	32.0 ± 6.3	26.8 ± 6.1	32.0 ± 6.4	26.5 ± 5.7	0.0 ± 4.0	−0.4 ± 3.6
Bent arm hang (s)	24.2 ± 36.3	9.5 ± 10.7	21.0 ± 29.8	10.5 ± 14.8	−0.7 ± 13.9	0.0 ± 16.3
4 × 10 m shuttle run (s)	11.7 ± 1.1	12.5 ± 1.1	11.8 ± 1.2	12.5 ± 1.0	0.1 ± 0.7	0.1 ± 0.8
20-m shuttle run (stages)	6.4 ± 2.3	4.1 ± 1.9	6.2 ± 2.5	4.0 ± 2.0	−0.1 ± 1.5	0.0 ± 1.1

[a]The average of right and left side scores is shown in the table and was used for the analyses. Both significant systematic bias and sex differences (*) in the studied physical fitness tests were analysed by one-way ANOVA, with sex as a fixed factor and inter-trial difference as the dependent variable. No significant systemic bias was found.

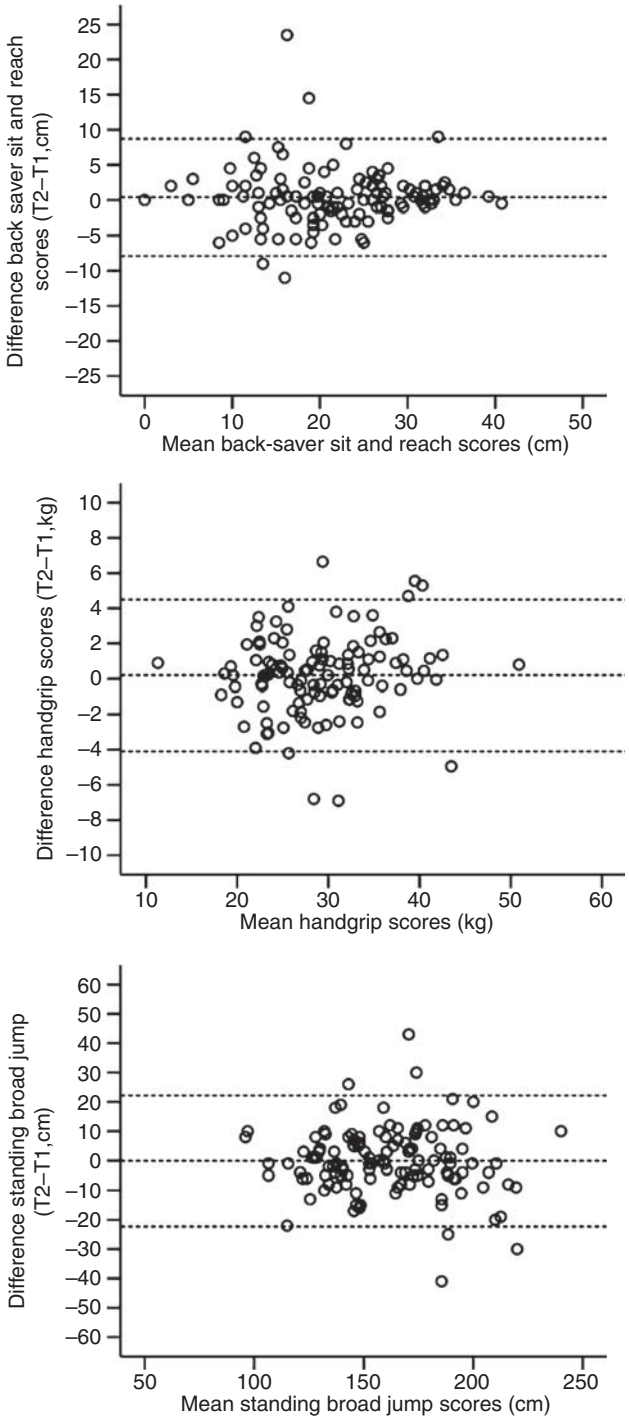

Figure 1 Bland–Altman plot of the back-saver sit and reach, handgrip and standing broad jump tests in adolescents. The central dotted line represents the mean differences between the second trial (T2) and the first trial (T1); the upper and lower dotted lines represent the upper and lower 95% limits of agreement (mean differences ± 1.96 s.d. of the differences), respectively.

Several reviews have proposed the Bland–Altman approach as an appropriate descriptive method for a meaningful and useful interpretation of reliability.[7–9] According to the review

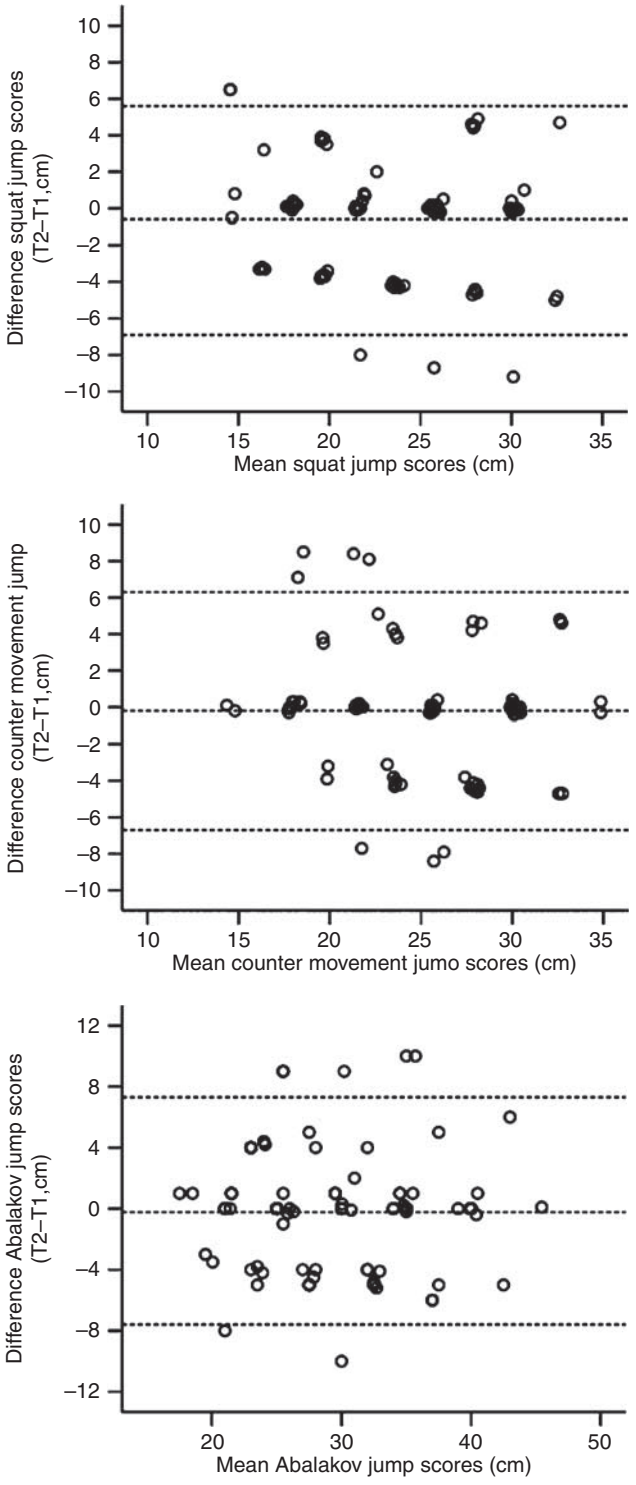

Figure 2 Bland–Altman plot of the Bosco jumps, that is, squat jump, counter movement jump and Abalakov jump in adolescents. The central dotted line represents the mean differences between the second trial (T2) and the first trial (T1); the upper and lower dotted lines represent the upper and lower 95% limits of agreement (mean differences ± 1.96 s.d. of the differences), respectively.

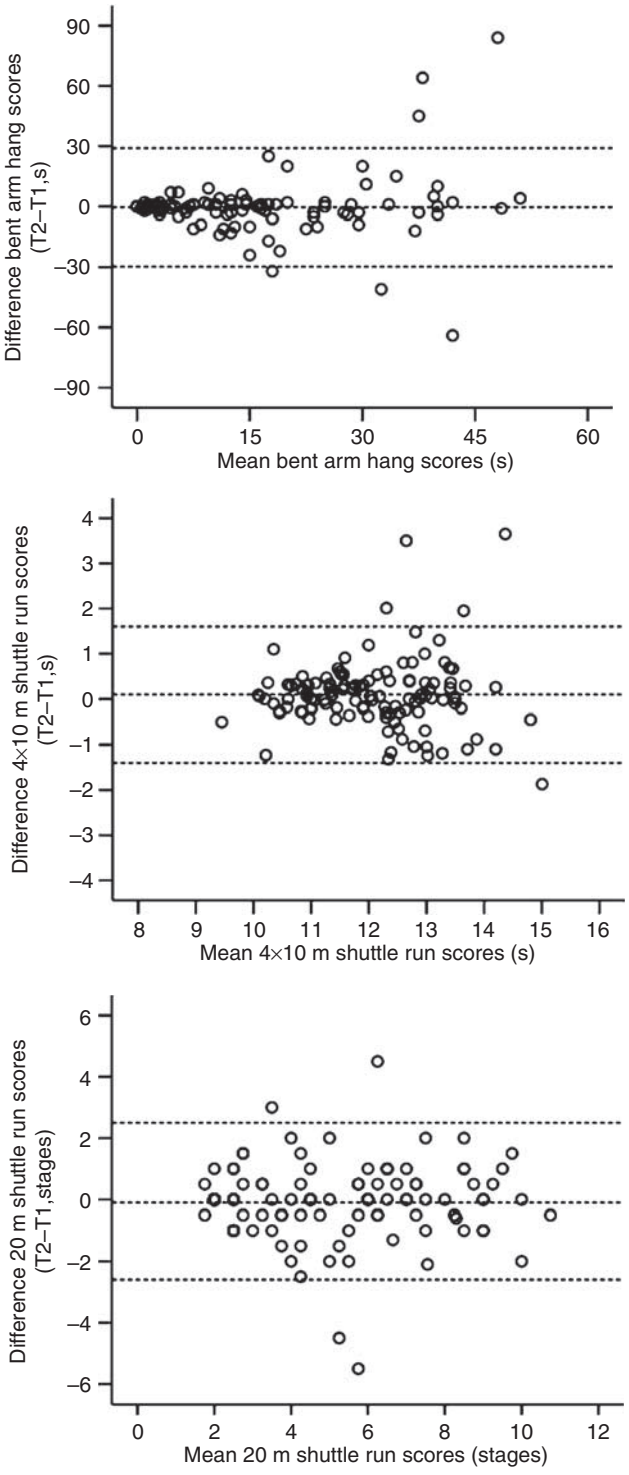

Figure 3 Bland–Altman plot of the bent arm hang, 4×10 m shuttle run and 20-m shuttle run tests in adolescents. The central dotted line represents the mean differences between the second trial (T2) and the first trial (T1); the upper and lower dotted lines represent the upper and lower 95% limits of agreement (mean differences ± 1.96 s.d. of the differences), respectively.

performed, only two (9%) of the physical fitness reliability studies carried out in healthy young people used the Bland–Altman approach.[10,35] When the search was extended to healthy or unhealthy people of any age, only eight additional studies using the Bland–Altman approach were found.[36–43] Given that physical fitness can already be considered a marker of health in this period of life,[3,6,44] information about the reliability of health-related physical fitness tests in young people is of interest. In addition, the review also shows that both cardiorespiratory fitness and muscular strength were the most studied physical fitness qualities (in 41 and 50%, respectively, of the reviewed studies), whereas data about speed-agility, coordination and flexibility in young people are lacking (used in 4–14% of the reviewed studies).

Collectively, the review and the methodological discussion performed above suggest that methods such as correlation or regression have important limitations and are not useful enough for studying reliability. In addition, the decision about what is 'acceptable' agreement is a scientific judgement; statistics alone cannot answer this question, as measurements, which may be considered to agree well enough for one purpose may not agree well enough for another.[45] For instance, if blood glucose concentration is measured twice, minutes apart, the acceptable error will be much lower if handgrip strength is measured 2 weeks apart.

Physical fitness reliability analysis

One of the main hypotheses tested in this study was whether a learning effect (positive systematic bias) exists among the physical fitness tests studied when repeated measurements are performed. Li *et al.*[10] examined the reliability of a 6-min walk test in adolescents. They found a bias of 15 m (95% limits of agreement: −35 to 65), whereas no significant difference was found between both measurements. Johnston *et al.*[35] studied the test-retest reliability of several physiological variables during a maximal cardiopulmonary exercise test in children. The peak oxygen consumption showed a bias of 1.4 ml kg^{-1} min^{-1} (95% limits of agreement: −3 to 5 ml kg^{-1} min^{-1}), but no significant difference was found between test and retest scores. In our study, the bias for the physical fitness variables studied in the study sample was close to 0 in most of the tests. The results suggest that neither learning nor fatigue (negative systematic bias) effects occurred when physical fitness was assessed with the tests used in this study, on a test-retest basis, in adolescents.

Results from the heteroscedasticity analyses and Bland–Altman plots indicate that the better (longer time) the performance in the bent arm hang test, the worse the degree of the agreement, whereas the better the performance in the back-saver sit and reach tests (further reach), the better the degree of the agreement.

In addition, the reliability of the physical fitness tests analysed is similar between male and female adolescents. This result is in accordance with data reported in men and

Table 2 Review of fitness reliability studies ($n = 22$) published since 1990 in healthy young people

Author	*Participants*	*Age (years)*	*Design*	*Fitness quality*	*Test*	*Statistical methods*
Erbaugh[15]	Boys = 13 Girls = 13	8.3 ± 1	Test-retest	Cardiorespiratory fitness, muscular endurance and flexibility	9-min run test, sit-up, modified pull-up, sit-and-reach	ANOVA, interclass correlation, ICC
Cotten[16]	Boys = 171 Girls = 192	5–12	Test-retest	Muscular endurance	Modified pull-up	ANOVA, ICC
Atwater *et al.*[17]	Boys and Girls = 24	4–9	Inter-rater and test-retest	Balance	One-leg balance and balance on a tiltboard	Wilcoxon matched pairs signed-rank tests, Spearman's rank-order correlation coefficients, magnitude of difference
Engelman and Morrow[18]	Boys = 242 Girls = 228	7–11	Test-retest	Muscular endurance	Pull-up (traditional and modified)	Two-way (alpha) reliability model
Kollath *et al.*[19]	Boys and Girls = 105	14–15	Test-retest	Muscular endurance	Modified pull-up	ANOVA, ICC, proportion of agreement, kappa statistic
Rikli *et al.*[20]	Boys = 177 Girls = 185	5–9	Test-retest	Cardiorespiratory fitness	1-mile run test	ANOVA, ICC
Liu *et al.*[21]	Boys = 12 Girls = 8	12–15	Test-retest	Cardiorespiratory fitness	20-m shuttle run test	ANOVA, ICC
Pate *et al.*[22]	Boys = 38 Girls = 56	9–10	Test-retest	Muscular endurance	Pull-up (original and two modified versions), flexed arm hang, push-up	ANOVA, ICC
Mc Manis and Wuest[23]	Boys = 57 Girls = 43	9 and 15	Test-retest	Muscular endurance	Modified push-up	ANOVA, ICC
Patterson *et al.*[24]	Boys = 42 Girls = 46	11–15	Test-retest	Flexibility	Back-saver sit and reach	ANOVA, ICC
Mahar *et al.*[25]	Boys = 137 Girls = 104	10–11	Test-retest	Cardiorespiratory fitness	PACER (adapted from 20-m shuttle run test)	ANOVA, ICC
Anderson *et al.*[26]	Boys = 107 Girls = 129	6–10	Test-retest	Muscular endurance	Curl-up	ANOVA, ICC
Patterson *et al.*[27]	Boys = 43 Girls = 45	11–15	Test-retest	Muscular strength, flexibility	Trunk lift	ANOVA, ICC, proportion of agreement, kappa statistic
McSwegin *et al.*[28]	Boys and Girls = 21	14–18	Test-retest	Cardiorespiratory fitness	1-mile walk test	Intraclass stability reliability coefficient
McManis *et al.*[29]	Boys and Girls = 310	Elementary and high school age	Test-retest	Muscular endurance	Push-up	Stability reliability coefficient
Figueroa-Colon *et al.*[30]	Girls = 61	7.3 ± 1.3	Test-retest	Cardiorespiratory fitness	Treadmill submaximal and maximal protocols	CV
Patterson *et al.*[31]	Boys = 36 Girls = 48	10–12	Inter-rater	Muscular endurance	Curl-up	ANOVA, ICC, proportion of agreement, kappa statistic
Tong *et al.*[32]	Boys = 14 Girls = 31	17	Test-retest	Cardiorespiratory fitness	5-min running field test	ANOVA, Pearson correlation coefficient (r), intraclass reliability coefficient, CV
Romain and Mahar[33]	Boys = 30 Girls = 32	11.4 ± 0.9	Test-retest	Muscular endurance	90° push-up, modified pull-up	ANOVA, ICC, proportion of agreement, kappa statistic
Alricsson *et al.*[34]	Boys = 8 Girls = 3	11	Test-retest	Speed and agility	Slalom test (speed) and hurdle test (agility)	ANOVA, ICC, CV
Li *et al.*[10]	Boys = 23 Girls = 29	14.2 ± 1.2	Test-retest	Cardiorespiratory fitness	6-min walk test	ICC, Bland–Altman limits of agreement
Johnston *et al.*[35]	Boys = 6 Girls = 3	8–11	Test-retest	Cardiorespiratory fitness	Treadmill continuous, incremental protocol	Wilcoxon matched pairs signed-rank tests, ICC, CV, Bland–Altman limits of agreement

Abbreviations: ANOVA, analysis of variance; CV, coefficient of variation; ICC, intraclass correlation coefficient.

women who performed a cardiopulmonary test in laboratory conditions.[39]

The wide variety of physical fitness tests examined in this study, the relatively large number of participants involved in the study and the use of adolescents from 10 European cities are the notable strengths of this study.

In conclusion, our study provides reference values for reliability of a wide set of physical fitness tests in European adolescents. Neither a learning nor a fatigue effect was found for any of the physical fitness tests when repeated. The results also suggest that reliability did not differ between male and female adolescents. Collectively, it can be stated

that the reliability of the physical fitness tests examined in this study is acceptable. The data provided contribute to a better understanding of physical fitness assessment in young people.

Acknowledgements

The HELENA Study was carried out with the financial support of the European Community Sixth RTD Framework Programme (Contract FOOD-CT-2005-007034). It is also being supported by grants from CSD in Spain (109/UPB31/03 and 13/UPB20/04), the Spanish Ministry of Education (EX-2007-1124; AP-2004-2745; AP2005-4358) and the ALPHA study, a European Union-funded study, in the framework of the Public Health Programme (Ref: 2006120). The researchers from the University of Zaragoza, Spain (GVR, JPRL) are complementarily supported by FUNDACIÓN MAPFRE (Spain). The content of this paper reflects only the authors' views, and the European Community is not liable for any use that may be made of the information contained therein. Finally, we acknowledge all participating children and adolescents, as well as their parents and teachers for their collaboration. We also acknowledge our staff members for their efforts and great enthusiasm during the fieldwork. The authors thank Professor Olle Carlsson for his assistance with the statistical analysis of the data.

Conflict of interest

The authors state no conflict of interest.

References

1 Myers J, Prakash M, Froelicher V, Do D, Partington S, Atwood JE. Exercise capacity and mortality among men referred for exercise testing. *N Engl J Med* 2002; **346**: 793–801.
2 Gulati M, Pandey DK, Arnsdorf MF, Lauderdale DS, Thisted RA, Wicklund RH *et al*. Exercise capacity and the risk of death in women: the St James Women Take Heart Project. *Circulation* 2003; **108**: 1554–1559.
3 Carnethon MR, Gulati M, Greenland P. Prevalence and cardiovascular disease correlates of low cardiorespiratory fitness in adolescents and adults. *JAMA* 2005; **294**: 2981–2988.
4 Andersen LB, Harro M, Sardinha LB, Froberg K, Ekelund U, Brage S *et al*. Physical activity and clustered cardiovascular risk in children: a cross-sectional study (The European Youth Heart Study). *Lancet* 2006; **368**: 299–304.
5 Moreno LA, González-Gross M, Kersting M, Molnár D, de Henauw S, Beghin L *et al*. Healthy lifestyle in Europe by nutrition in adolescence. The HELENA Study. *Public Health Nutr* 2008; **11**: 288–299.
6 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**: 269–277.
7 Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998; **26**: 217–238.
8 Rothwell PM. Analysis of agreement between measurements of continuous variables: general principles and lessons from studies of imaging of carotid stenosis. *J Neurol* 2000; **247**: 825–834.
9 Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? *Physiotherapy* 2000; **86**: 94–99.
10 Li AM, Yin J, Yu CC, Tsang T, So HK, Wong E *et al*. The six-minute walk test in healthy children: reliability and validity. *Eur Respir J* 2005; **25**: 1057–1060.
11 Ruiz JR, Espana-Romero V, Ortega FB, Sjostrom M, Castillo MJ, Gutierrez A. Hand span influences optimal grip span in male and female teenagers. *J Hand Surg [Am]* 2006; **31**: 1367–1372.
12 Leger LA, Mercier D, Gadoury C, Lambert J. The multistage 20 metre shuttle run test for aerobic fitness. *J Sports Sci* 1988; **6**: 93–101.
13 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**: 307–317.
14 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–310.
15 Erbaugh SJ. Reliability of physical fitness tests administered to young children. *Percept Mot Skills* 1990; **71**: 1123–1128.
16 Cotten DJ. An analysis of the NCYFS II Modified Pull-up Test. *Res Q Exerc Sport* 1990; **61**: 272–274.
17 Atwater SW, Crowe TK, Deitz JC, Richardson PK. Interrater and test-retest reliability of two pediatric balance tests. *Phys Ther* 1990; **70**: 79–87.
18 Engelman ME, Morrow Jr JR. Reliability and skinfold correlates for traditional and modified pull-ups in children grades 3–5. *Res Q Exerc Sport* 1991; **62**: 88–91.
19 Kollath JA, Safrit MJ, Zhu W, Gao LG. Measurement errors in modified pull-ups testing. *Res Q Exerc Sport* 1991; **62**: 432–435.
20 Rikli RE, Petray C, Baumgartner TA. The reliability of distance run tests for children in grades K-4. *Res Q Exerc Sport* 1992; **63**: 270–276.
21 Liu NY, Plowman SA, Looney MA. The reliability and validity of the 20-meter shuttle test in American students 12–15 years old. *Res Q Exerc Sport* 1992; **63**: 360–365.
22 Pate RR, Burgess ML, Woods JA, Ross JG, Baumgartner T. Validity of field tests of upper body muscular strength. *Res Q Exerc Sport* 1993; **64**: 17–24.
23 McManis BG, Wuest DA. Stability reliability of the modified push-up in children. *Res Q Exerc Sport* 1994; **65** (Suppl): A58–A59 (abstract).
24 Patterson P, Wiksten DL, Ray L, Flanders C, Sanphy D. The validity and reliability of the back saver sit-and-reach test in middle school girls and boys. *Res Q Exerc Sport* 1996; **67**: 448–451.
25 Mahar MT, Rowe DA, Parker CR, Mahar FJ, Dawson DM, Holt JE. Criterion-referenced and norm-referenced agreement between the mile run/walk and PACER. *Meas Phys Educ Exerc Sci* 1997; **1**: 245–258.
26 Anderson EA, Zhang JJ, Rudisill ME, Gaa J. Validity and reliability of a timed curl-up test: development of a parallel form for the FITNESSGRAM abdominal strength test. *Res Q Exerc Sport* 1997; **68** (Suppl): A-51.
27 Patterson P, Rethwisch N, Wiksten D. Reliability of the trunk lift in high school boys and girls. *Meas Phys Educ Exerc Sci* 1997; **1**: 145–151.
28 McSwegin PJ, Plowman SA, Wolff GM, Guttenberg GL. The validity of a one-mile walk test for high school age individuals. *Meas Phys Educ Exerc Sci* 1998; **2**: 47–63.
29 McManis BG, Baumgartner TA, West DA. Objectivity and reliability of the 90° pushup test. *Meas Phys Educ Exerc Sci* 2000; **4**: 57–67.
30 Figueroa-Colon R, Hunter GR, Mayo MS, Aldridge RA, Goran MI, Weinsier RL. Reliability of treadmill measures and criteria to determine VO2max in prepubertal girls. *Med Sci Sports Exerc* 2000; **32**: 865–869.

31 Patterson P, Bennington J, De La Rosa T. Psychometric properties of child- and teacher-reported curl-up scores in children aged 10–12 years. *Res Q Exerc Sport* 2001; **72**: 117–124.
32 Tong TK, Fu FH, Chow BC. Reliability of a 5-min running field test and its accuracy in VO2max evaluation. *J Sports Med Phys Fitness* 2001; **41**: 318–323.
33 Romain BS, Mahar MT. Norm-referenced and criterion-referenced reliability of the push-up and modified pull-up. *Meas Phys Educ Exerc Sci* 2001; **5**: 67–80.
34 Alricsson M, Harms-Ringdahl K, Werner S. Reliability of sports related functional tests with emphasis on speed and agility in young athletes. *Scand J Med Sci Sports* 2001; **11**: 229–232.
35 Johnston KN, Jenkins SC, Stick SM. Repeatability of peak oxygen uptake in children who are healthy. *Pediatr Phys Ther* 2005; **17**: 11–17.
36 de Greef MH, Sprenger SR, Elzenga CT, Popkema DY, Bennekers JH, Niemeijer MG *et al*. Reliability and validity of a twelve-minute walking test for coronary heart disease patients. *Percept Mot Skills* 2005; **100**: 567–575.
37 Taylor S, Frost H, Taylor A, Barker K. Reliability and responsiveness of the shuttle walking test in patients with chronic low back pain. *Physiother Res Int* 2001; **6**: 170–178.
38 Buckley JP, Sim J, Eston RG, Hession R, Fox R. Reliability and validity of measures taken during the Chester step test to predict aerobic power and to prescribe aerobic exercise. *Br J Sports Med* 2004; **38**: 197–205.
39 Bingisser R, Kaplan V, Scherer T, Russi EW, Bloch KE. Effect of training on repeatability of cardiopulmonary exercise performance in normal men and women. *Med Sci Sports Exerc* 1997; **29**: 1499–1504.
40 Lamb KL, Eston RG, Corns D. Reliability of ratings of perceived exertion during progressive treadmill exercise. *Br J Sports Med* 1999; **33**: 336–339.
41 van 't Hul A, Gosselink R, Kwakkel G. Constant-load cycle endurance performance: test-retest reliability and validity in patients with COPD. *J Cardiopulm Rehabil* 2003; **23**: 143–150.
42 Balfour-Lynn IM, Prasad SA, Laverty A, Whitehead BF, Dinwiddie R. A step in the right direction: assessing exercise tolerance in cystic fibrosis. *Pediatr Pulmonol* 1998; **25**: 278–284.
43 Wallman K, Goodman C, Morton A, Grove R, Dawson B. Test-retest reliability of the aerobic power index component of the tri-level fitness profile in a sedentary population. *J Sci Med Sport* 2003; **6**: 443–454.
44 Ruiz JR, Ortega FB, Meusel D, Harro M, Oja P, Sjöström M. Cardiorespiratory fitness is associated with features of metabolic risk factors in children. Should cardiorespiratory fitness be assessed in a European health monitoring system? The European Youth Heart Study. *J Public Health* 2006; **14**: 94–102.
45 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–160.