Journal of Science and Medicine in Sport 26 (2023) 553–560

Contents lists available at ScienceDirect

Journal of Science and Medicine in Sport

journal homepage: www.elsevier.com/locate/jsams

Original research

Time measurement validity and reliability of the 4 × 10-m shuttle run test in adult population: The ADULT-FIT project

José Castro-Piñero [a,b], Pedro Aragón-Aragón [a], Carolina Cruz-León [a,b], José Jiménez-Iglesias [a,b,c], Daniel Camiletti-Moirón [a,b,*], Rocío Izquierdo-Gómez [a,b], Magdalena Cuenca-García [a,b]

[a] GALENO research group, Department of Physical Education, Faculty of Education Sciences, University of Cadiz, Spain
[b] Instituto de Investigación e Innovación Biomédica de Cádiz (INiBICA), Spain
[c] Sport Science Department Cádiz C.F., Cádiz C.F., Spain

ARTICLE INFO

Article history:
Received 15 May 2023
Received in revised form 18 July 2023
Accepted 10 August 2023
Available online 23 August 2023

Keywords:
Physical fitness
Motor fitness
Validation
Reproducibility
Field-based test
Adults

ABSTRACT

Objectives: The purpose of this study was to analyze the time measurement validity and reliability (between raters and test–retest) of the 4 × 10-m shuttle run test to assess motor fitness in adults, according to gender, age, and physical activity levels.
Design: Cross-sectional. A total of 230 adults (86 women) aged 18–64 years participated in the study.
Methods: The time taken to complete the 4 × 10-m shuttle run test was recorded simultaneously by a trained and an untrained rater (inter-rater reliability) and by photoelectric cells (time measurement validity). 48–72 h later, the test was repeated under the same conditions (test–retest reliability).
Results: The systematic error for trained rater vs. photocell was close to zero (0.0125, $p < 0.01$), with an effect size of 0.006; and for both, untrained rater vs. photocell and trained rater vs. untrained rater was ~0.2 s ($p < 0.001$) with an effect size of 0.09. For the test–retest reliability, the systematic error was 0.05 s ($p < 0.001$), with an effect size of 0.26, the intraclass correlation coefficient was 0.998 and the coefficient of variation reported a variability of 0.73 %. Results were not influenced by gender and age, while these improved for active vs. non-active participants.
Conclusions: Findings indicate that measurements with trained raters are a valid and reliable method for assessing the 4 × 10-m shuttle run test in adults. It is highly recommended that raters be trained to minimize the measurement error.

Practical implications

- Researchers, coaches and health professionals can correctly measure the 4 × 10-m shuttle run test using a stopwatch, reducing the error when time is recorded by a trained rater.
- Untrained raters could be used in population-based studies as long as the presence of a greater error in the measurements is assumed.
- The time measurement validity and the reliability of the 4 × 10-m shuttle run test improve when the sample corresponds to physically active adults, regardless of gender and age groups.
- The 4 × 10-m shuttle run test is useful, practical, as well as easy to perform, efficient in time and with low personal cost to assess motor fitness in adult population.

1. Introduction

Physical fitness is a powerful marker of general health in adults (18–64 years), especially cardiorespiratory fitness and muscular fitness. Cardiorespiratory fitness is an important predictor of all-cause mortality[1] and is inversely related to cardiovascular diseases and diabetes,[2] and some types of cancer.[3] In addition, it is associated with improved well-being and cognitive function and reduces the risk of Alzheimer's.[4] Likewise, muscular fitness is a relevant protector against all-cause mortality and reduces the prevalence and incidence of metabolic syndrome.[5,6]

Recently, a systematic review has shown the predictive validity of motor fitness (i.e., speed, agility, coordination, and balance) in adults.[7] It is justified that the motor fitness assessment is a key tool for diagnosing the health state in adults, given its close relationship with cardiovascular disease, fall and the risk of falls, cognitive impairment, mobility disability, hospitalization, disability in instrumental activities of daily living, and all-cause mortality.

* Corresponding author.
E-mail address: daniel.camiletti@uca.es (D. Camiletti-Moirón)
@DCamiletti.

https://doi.org/10.1016/j.jsams.2023.08.176

The validity and reliability of field-based cardiorespiratory and muscular fitness tests have been widely assessed in adults, however little evidence exists in this regard with field-based motor fitness tests, requiring more studies, especially high-quality ones.[8,9]

The 4 × 10-m shuttle run test is widely used in preschoolers and children and adolescents to assess motor fitness. In fact, this test has been included in several field-based physical fitness test batteries in preschoolers (i.e., PREFIT battery),[10] and in children and adolescents (i.e., ALPHA-Fitness battery),[11] showing a good validity and reliability for the young population.[12,13]

Although, the 4 × 10-m shuttle run test has been used in the scientific literature in order to assess motor fitness as a marker of health in healthy adults[14,15] or specific populations, such as adults with ankle instability[16] and Parkinson's patients,[17] and it has also been proposed to assess the beneficial effect of training on physiological parameters, such as insulin-like growth factor 1 (IGF-1), testosterone[18] and anthropometric measurements,[19] there is still no evidence of its validity and reliability in the healthy adult population.

For this reason, although further studies, especially longitudinal ones, would be necessary, the 4 × 10-m shuttle run test could be a useful and effective alternative to assess health-related motor fitness in adults. Thus, it would be desirable to establish its time measurement validity and reliability (between raters and test–retest) in a full age range of the adult population (i.e., 18–64 years), which has been confirmed in adolescents,[13] considering gender and physical activity levels. However, the level of evidence of time measurement validity and inter-rater reliability of the 4 × 10-m shuttle run test is still limited, requiring more studies to establish strong evidence.[8]

Therefore, the aim of this study is to analyze the time measurement validity and the reliability (between raters and test–retest) of the 4 × 10-m shuttle run test to assess motor fitness (i.e., speed–agility) in adults (18–64 years), according to gender, age, and physical activity levels.

We hypothesize that the 4 × 10-m shuttle run test is valid and reliable (between raters and test–retest), especially when the rater is trained, being important to know whether the gender, age or level of physical activity can play a determining role.

2. Methods

The data from this study are part of the research project "the ADULT-FIT study" (DEP2017-88043-R), whose main purpose was to design a field-based physical fitness-test battery related to health based on their criterion-validity, predictive validity, reliability, feasibility, and safety for use in adults. A total of 230 adults (86 females) from Cadiz (Spain), participated in this study. The sample was distributed by gender, age (18–34 years, 35–49 years, and 50–64 years), and physical activity levels (non-active and active).

All participants signed an informed consent on the first day of measurement in the laboratory prior to performing the tests. The study was approved by the Committee for Research of Cadiz, Spain.

The participants were evaluated in two different sessions carried out over 48–72 h. During the first session, the 4 × 10-m shuttle run test was performed to assess the time measurement validity (*manual* vs. *automatic* timing) and the inter-rater reliability (i.e., *homogeneity*, *trained* vs. *untrained rater*). In the second session (48–72 h later), the 4 × 10-m shuttle run test protocol was carried out again in the same conditions as the previous session, to analyze the test–retest reliability.

Participants were asked, how much time do you practice physical activity/exercise or some sport, of at least moderate intensity, per day?, and were classified as active/non-active when following/not following the World Health Organization recommendations for adults.[20]

Two parallel lines were marked on the ground 10 m apart. Participants ran back and forth as quickly as possible, crossing each line with both feet each time, covering a total distance of 40 m (4 × 10-m). Every time the participant crossed any of the lines, the participant picked up (the first time) a sponge, which was previously placed behind the lines, or exchanged (the second and third times) a sponge behind the lines.[11] The time was measured using photoelectric cells (Photocells Chronojump races, Chronojump-Boscosystem, Madrid, Spain) (1st session) and a manual stopwatch (Casio HS-EV-1RET Digital, Casio, Tokyo, Japan) (1st and 2nd session), and stopped when the participant crossed the finish line with one foot, having the last sponge in one of his/her hands, without the need to place it on the ground.

The test was performed twice with at least 2 min of rest between attempts, in each session. The participants were verbally motivated during the tests, and the best time nearest tenth of a second, obtained in each session, was later used for analysis.

2.1. Selection of raters (i.e., timekeepers)

There were two raters, one trained and one untrained. The trained rater (R1) had previously participated in the training process and in the pilot study of the ADULT-FIT project. In addition, R1 participated in the time measurement validity test, the inter-rater reliability test and the test–retest. The untrained rater (R2) had not attended the training process or the pilot study of the project, participating in the time measurement validity and in the inter-rater reliability tests.

2.2. Time measurement validity study

In the first session, two attempts were performed in the 4 × 10-m shuttle run test, and were automatically assessed with photoelectric cells, apart from R1 and R2. In fitness tests and sports events, photoelectric cells are used as criterion reference to assess the time taken to cross two separate lines.[21] The photoelectric cell timer was automatically activated when the participant crossed the first cell and stopped when the participant crossed the last cell at the finish line. For this test, the photoelectric cells were placed at the start and finish line (it was the same line), where the raters were also located, having an activation range from 5 to 35 cm above the ground.

2.3. Inter-rater reliability (homogeneity) study

R1 and R2 simultaneously measured the time required to complete the test in the first session with a manual stopwatch. The raters started the stopwatch when the participant crossed the starting line and stopped it when the participant crossed the finish line.

2.4. Test–retest reliability study

The records taken by R1 in the first session (test, T1) and in the second session (retest, T2) were used. As we pointed out before, the second session was carried out 48–72 h after the first session, and in both sessions the same protocol and conditions were developed.

The normality test was carried out by the Kolmogorov–Smirnov test, to know whether variables followed a normal distribution. Since the sample followed a normal distribution, each variable was presented as mean (standard deviation). An analysis of variance (ANOVA) was used to check whether there were significant differences between the gender, age and physical activity level groups.

Concordance between the manual stopwatch (i.e., R1 and R2) and photoelectric cells, and the relationship between the times measured by R1 vs. R2 (i.e., inter-rater reliability) were analyzed graphically following the Bland–Altman method.[22] The limits of agreement (LoA) were established as systematic error (mean difference) ± 1.96 standard deviations of the difference. Differences were plotted against the gold standard (i.e., photoelectric cells) rather than the mean value because this was expected to be closer to the "true value" than the mean.[23] In addition, in both cases, the presence of systematic errors was analyzed by repeated measured of ANOVA. The factor pairs included in the analysis were R1 vs. R2, R1 vs. photoelectric cells, and R2 vs. photoelectric cells.

Table 1
Descriptive characteristics of the sample.

	All (n = 230)	Gender		Age groups			Physical activity levels	
		Male (n = 144)	Female (n = 86)	18–34 years (n = 133)	35–49 years (n = 51)	50–64 years (n = 46)	Active (n = 110)	Non-active (n = 120)
Age (years)	35.40 (15.13)	29.93 (13.26)***	44.56 (13.63)	24.68 (9.03)XXX	44.16 (3.80)YYY	56.70 (4.48)ZZZ	22.29 (4.67)▲▲▲	47.42 (10.80)
Measurement method								
Photoelectric cell (s)	11.95 (2.13)	10.84 (1.41)***	13.79 (1.83)	10.68 (1.54)XXX	13.20 (1.50)	14.22 (1.33)ZZZ	10.08 (0.47)▲▲▲	13.66 (1.53)
Trained (R1) (s)	11.96 (2.13)	10.85 (1.41)***	13.81 (1.83)	10.69 (1.54)XXX	13.21 (1.50)	14.24 (1.33)ZZZ	10.09 (0.46)▲▲▲	13.68 (1.53)
Untrained rater (R2) (s)	12.14 (2.18)	11.04 (1.49)***	13.99 (1.88)	10.84 (1.58)XXX	13.47 (1.56)	14.44 (1.34)ZZZ	10.23 (0.50)▲▲▲	13.90 (1.57)
Test–retest								
Test (s)	12.36 (2.08)	11.32 (1.35)***	14.09 (1.93)	11.17 (1.51)XXX	13.48 (1.54)	14.55 (1.46)ZZZ	10.64 (0.64)▲▲▲	13.93 (1.65)
Retest (s)	12.30 (2.02)	11.28 (1.31)***	14.01 (1.85)	11.15 (1.46)XXX	13.41 (1.57)	14.42 (1.38)ZZZ	10.63 (0.63)▲▲▲	13.83 (1.60)

Values are presented as mean (standard deviation).
Differences between gender: ***$p < 0.001$; differences between age groups 18–34 years & 35–49 years: XXX$p < 0.001$; differences between age groups 35–49 years & 50–64 years: YYY$p < 0.001$; differences between age groups 18–34 years & 50–64 years: ZZZ$p < 0.001$; differences between physical activity levels: ▲▲▲$p < 0.001$.

For test–retest reliability, the paired *t*-test, the intraclass correlation coefficient (ICC) with 95 % confident intervals (CI) and the Bland–Altman method were used. Moreover, we also examined the differences between T1 and T2 using different error measures. Generally, the lower the error value, the lower the dispersion between test and retest measurements.

The sum of squared errors (SSE) was calculated as follows:

$$SSE = \sum_{i=1}^{N} (yi - \hat{y})^2$$

where N is the cases to evaluate the error measurements, $\hat{y}$ is the T2, and y is the T1. The mean sum of squared errors (MSE):

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (yi - \hat{y})^2$$

The root mean sum of squared errors (RMSE) was calculated by converting MSE into domain units by taking the root square:

$$RMSE = \sqrt{MSE}$$

The percentage error was calculated as follows:

$$\%Error = \frac{RMSE}{y_{max} - y_{min}} \times 100$$

The standard error of measurement (SEM) is expressed in standardized values, as a percentage of the mean value of the measurements. The SEM quantifies the precision of individual scores in a test, and it is not influenced by variability among individuals (i.e., is considered a fixed characteristic of any measure, regardless of the sample of participants under investigation). A value ≤ 15 % is considered acceptable.[24]

The %SEM was calculated as follows: %SEM = mean of the difference scores between 2 trials × 100 / mean of the first trial.

To estimate the smallest change in score that indicates a "real change" in 90 % of participants, the minimal detectable change (MDC_{90})[25] was calculated:

$$MDC_{90} = SEM \times \sqrt{2} \times 1.65$$

The coefficient of variation (CV) was calculated as follow:

$$\%CV = \frac{\delta}{\overline{X}} \times 100$$

The CV method provides useful information in the presence of heteroscedasticity (assumes that greatest T1 and T2 variation occurs in individuals scoring the highest values in the test). A CV ≤ 10 % was considered as acceptable reliability.[26]

The standard error of estimate (SEE) was calculated as follows:

$$SEE = SD\hat{y}\sqrt{\left(1 - R^2 y\hat{y}\right)}$$

Complementary, the presence of heteroscedasticity was analyzed using ANOVA, establishing the absolute difference as the dependent variable (negative values were multiplied by −1) and factor the quartiles of magnitude (in this case, the photoelectric cells).[27] As well as, after making each Bland–Altman plot, a linear regression was carried out to analyze the presence of proportional bias; and Cohen's *d* was calculated to determine the effect size.[28]

After carrying out the statistical analysis for the entire sample of participants, an analysis was performed by gender, age and physical activity level groups, applying the same statistical procedure for time measurement validity, inter-rater reliability and test–retest reliability.

All the analyses were performed using the Statistical Package for Social Sciences (IBM SPSS Statistics for Windows, version 26.0; Armonk, NY) and the level of significance was set at $p < 0.05$.

3. Results

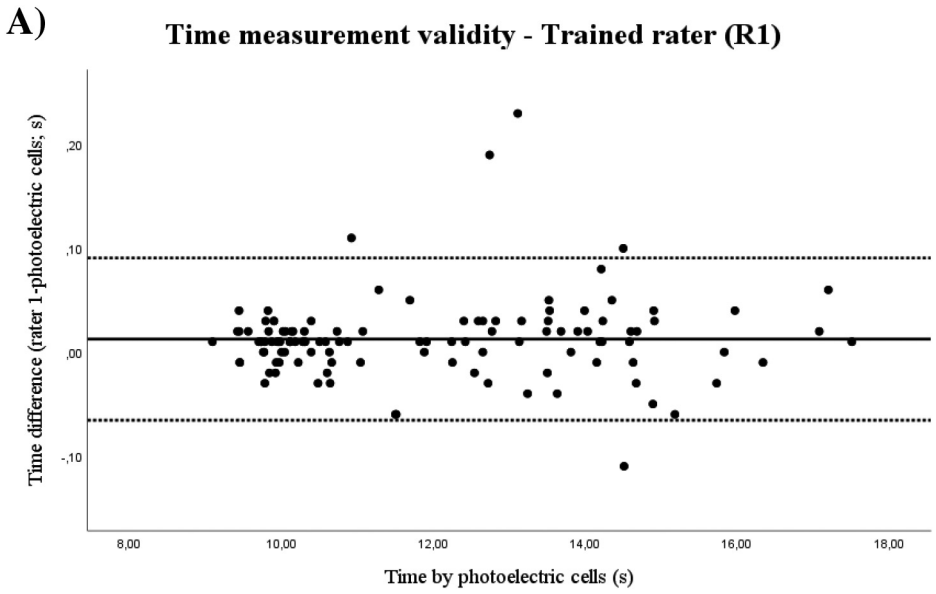
The descriptive characteristics of the participants by gender, age groups and physical activity levels are presented in Table 1. Males obtained significantly better times than females (all $p < 0.001$), regardless of the method and the measurement moment (test–retest). The 18–34 year old group presented significantly better times in the 4 × 10-m shuttle run test (all $p < 0.001$) compared to the 35–49 year old and 50–64 year old groups, regardless of the method and the measurement moment (test–retest). No significant differences were found between the age groups 35–49 years and 50–64 years in any of the variables analyzed. The active group showed lower times in the 4 × 10-m shuttle run test regardless of the method and the measurement moment (all $p < 0.001$), than the non-active group.

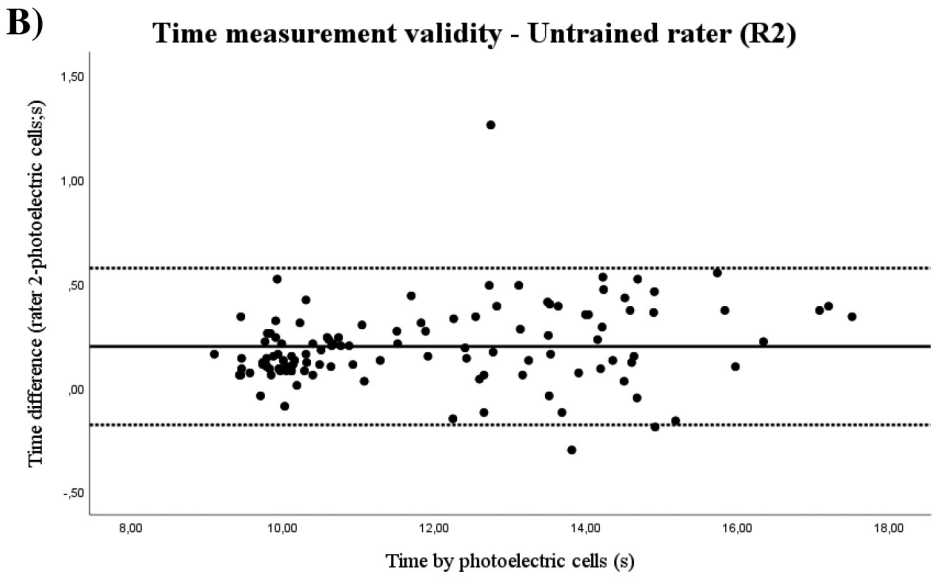
The time measurement validity of the 4 × 10-m shuttle run test is shown in Table 2. We found a systematic error of ~0.2 s for R2 ($p < 0.001$), and 0.01 s for R1 ($p < 0.01$). The effect size for R2 was 0.09 and for R1 was 0.006 (Table 2). Graphically, the LoA were established between 0.04 and −0.07 s for the R1 (Fig. 1A); and between 0.57 and −0.18 s for the R2 (Fig. 1B). Presence of heteroscedasticity ($p < 0.05$)

Table 2
Time measurement validity (trained and untrained raters vs. photoelectric cells) and inter-rater reliability (untrained vs. trained raters) of the 4 × 10-m shuttle run test.

	Systematic error (s)[a]	95 % limits of agreement[b]		Effect size (Cohen's *d*)
Time measurement validity				
Trained rater vs. photoelectric cells (s)	0.0125[c]	0.04	−0.07	0.006
Untrained rater vs. photoelectric cells (s)	0.1967[c]	0.57	−0.18	0.091
Inter-rater reliability				
Untrained vs. trained raters (s)	0.1842[c]	0.49	−0.18	0.085

[a] Systematic error, mean difference.
[b] 95 % limits of agreement, mean difference ± 1.96 SDs of the difference.
[c] $p < 0.001$, analyzed by repeated measured of ANOVA.


A)
Time measurement validity - Trained rater (R1)
Time difference (rater 1-photoelectric cells; s)
,20
,10
,00
-,10
8,00
10,00
12,00
14,00
16,00
18,00
Time by photoelectric cells (s)


B)
Time measurement validity - Untrained rater (R2)
Time difference (rater 2-photoelectric cells;s)
1,50
1,00
,50
,00
-,50
8,00
10,00
12,00
14,00
16,00
18,00
Time by photoelectric cells (s)

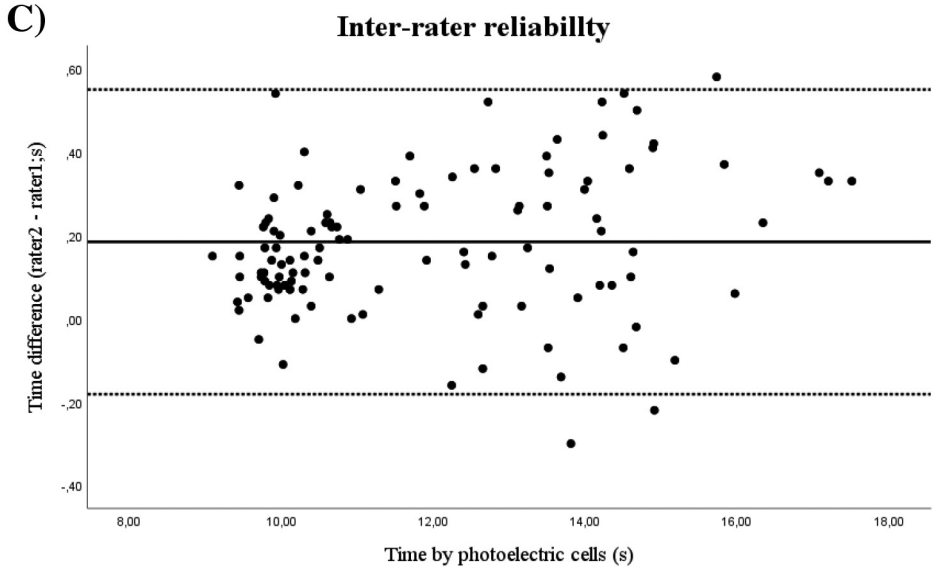

C)
Inter-rater reliabillty
Time difference (rater2 - rater1;s)
,60
,40
,20
,00
-,20
-,40
8,00
10,00
12,00
14,00
16,00
18,00
Time by photoelectric cells (s)

Table 3
Test–retest reliability of the 4 × 10-m shuttle run test.

	T1[a]	T2[a]	Intertrial difference (T2 − T1)	p value	Cohen's *d*	ICC (95 % CI)[**]	SSE	MSE	RMSE	% error	% SEM	% CV	SEE
4 × 10-m shuttle run test (s)	12.36 ± 2.08	12.30 ± 2.02	−0.06 ± 0.18 Retest < test[b] (n = 130)	<0.001	0.026	0.998 (0.997–0.998)	8.08	0.04	0.19	2.34	0.43	0.73	0.17

ICC, intraclass correlation coefficients; CI, confident interval; SSE, sum of squared errors; MSE, mean sum of squared errors; RMSE, root mean sum of squared errors; % error, percentage error; % SEM, standard error of measurement; % CV, percentage coefficient of variation; SEE, standard error of estimate.
[a] T1 refers to test (trial 1) and T2 to retest (trial 2). T2 − T1 refers to retest (trial 2) minus test (trial 1). Values are displayed as mean ± SD.
[b] The time recorded by the test is higher than by the retest.
[**] ICC was significant at $p < 0.001$.

was observed, showing that the longer the time spent performing the test, the more the values move away from the mean difference. On the other hand, the linear regression model between R1 and the photoelectric cells did not show any proportional bias ($p = 0.351$, $R^2 = 0.004$), unlike R2 and the photoelectric that showed a positive proportional bias ($p < 0.01$, $R^2 = 0.047$). The fact that the longer the test lasted the greater the error was, was unexpected since the higher the speed there is usually a greater measurement error. More studies are needed to elucidate this finding.

No relevant differences were observed between genders in the time measurement validity (data not shown), unlike the physical activity level groups where there seemed to be important changes to take into account. There were also important changes in the age groups (18–34 years vs. 35–49 and 50–64 years), but we hypothesized that they were determined by the physical activity levels, since the group of 18–34 years was the active one and the rest of the participants corresponded to the non-active ones. Therefore, we focused on analyzing the time measurement validity and the reliability (between raters and test–retest) of the 4 × 10-m shuttle run test by physical activity level groups.

From the analysis by physical activity level groups, we observed that there was no systematic error of the active group for R1 ($p > 0.05$) while for R2 the systematic error was ~0.15 s ($p < 0.001$). For the non-active group, there was no systematic error of R1 ($p > 0.05$), while for R2 the systematic error was significantly higher (0.24; $p < 0.001$). The effect size for R2 was 0.3 for the active group and 0.15 for the non-active group (Supplementary Table 1). Graphically (Supplementary Fig. 1) we observed, that for R1 the LoA were established at 0.04 and −0.025 s for the active group, while for the non-active group the LoA were established between 0.12 and −0.08 s. For R2, the LoA for the non-active group were significantly higher ($p < 0.001$). The presence of heteroscedasticity disappeared when dividing the sample by active and non-active groups ($p > 0.05$). The linear regression model did not show any proportional bias ($p > 0.05$) for R1 or for R2, in both the active and non-active groups.

We found a systematic error between R2 and R1 of ~0.18 s ($p < 0.001$), with an effect size of 0.085 (Table 2). The reliability pattern between raters is shown graphically in Fig. 1C. The LoA were established between 0.49 and −0.18 s. Presence of heteroscedasticity ($p < 0.05$) was observed, indicating that the longer the time spent performing the test, the more the values move away from the mean difference. The inter-rater linear regression model showed the presence of a positive proportional bias ($p < 0.01$, $R^2 = 0.044$).

From the analysis by physical activity level groups, we found that the systematic error for the active group was 0.15 s ($p < 0.001$) and for the non-active group was 0.21 s ($p < 0.001$); with an effect size of 0.3 and 0.14, respectively (Supplementary Table 1). Graphically, the LoA were established between 0.36 and −0.065 s for the active group, and between 0.67 and −0.24 for the non-active group (Supplementary Fig. 1). The presence of heteroscedasticity also disappeared in this case when dividing the sample by groups ($p > 0.05$). In the linear regression model, we observed that for both, the active and non-active groups, there was no proportional bias ($p > 0.05$), unlike the entire group.

The statistical analysis applied to determine the test–retest reliability of the 4 × 10-m shuttle run test is presented in Table 3. The systematic error was 0.06 s ($p < 0.01$), with an effect size of 0.03, and 56 % of the times recorded in the T2 were lower than in the T1. The ICC showed excellent agreement (ICC; [95 % CI] = 0.998; [0.997–0.999]) between the recorded times. All the error measurements reported low values (%Error = 2.34; %SEM = 0.43; %CV = 0.73; SEE = 0.17). Finally, the MDC_{90} was around 0.00 for each measure, indicating that there has been no real change between T1 and T2.

The test–retest reliability patterns are shown graphically in Fig. 2. The LoA were established between 0.302 and −0.407 s. Presence of heteroscedasticity ($p < 0.001$) was observed, indicating that the longer the time spent performing the test, the more the values move away from the mean difference. The linear regression model showed the presence of a positive proportional bias ($p < 0.001$), although the model only predicts ~10.4 % of the values.

The analysis by physical activity level groups, reported that in the active group there were no significant differences between the test and the retest ($p > 0.05$), while in the non-active group they were −0.1 s ($p < 0.01$); with an effect size of 0.06 (Supplementary Table 2). The error measurements also presented low values in both groups. Graphically, we observed a LoA between 0.18 and −0.17 s for the active group, and a LoA between 0.36 and −0.55 s for the non-active group (Supplementary Fig. 2). Heteroscedasticity disappeared when we divided the sample by groups ($p > 0.05$). The linear regression model did not show the presence of any proportional bias in both active and non-active groups ($p > 0.05$).

4. Discussion

The aim of the present study was to analyze the time measurement validity and the reliability (between raters and test–retest) of the 4 × 10-m shuttle run test to assess motor fitness (i.e., speed–agility) in adults (18–64 years), according to gender, age, and physical activity levels.

The main findings were that the 4 × 10-m shuttle run test provides valid and reliable results when the test was measured by trained raters, unlike untrained raters, where the accuracy of the data recorded was significantly lower. The time measurement validity and the reliability of the test were not influenced by gender and age groups, while both parameters improved when the sample corresponds to physically active adults, regardless of the type of rater.

To perform speed and/or agility test assessment, the ideal scenario is to use a reference measurement method (i.e., gold standard), such as photoelectric cells. These are not usually accessible in most contexts outside the laboratory, or applicable in epidemiological studies, and its

Fig. 1. Bland–Altman plots for the 4 × 10-m shuttle run test time measurement validity, showing the mean difference between time measured by R1 (A) and R2 (B) and time measured by photoelectric cells vs. time measured by photoelectric cells. (C) Interrater reliability showing the mean difference between R2 and R1 vs. time measured by photoelectric cells for the 4 × 10-m shuttle run test. The central solid line represents the mean differences (systematic error). The upper and lower dotted lines represent the upper and lower 95 % limits of agreement (mean differences ± 1.96 SDs of the differences), respectively.

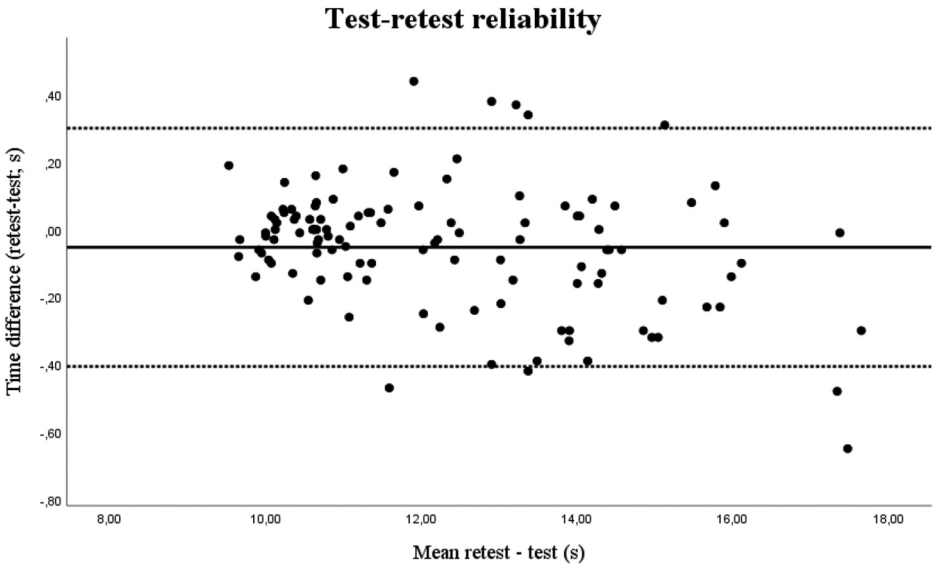


Fig. 2. Bland–Altman plot for the 4 × 10-m shuttle run test–retest reliability. The central solid line represents the mean difference between the retest and the test (systematic error). The upper and lower dotted lines represent the upper and lower 95 % limits of agreement (mean differences ± 1.96 SDs of the differences), respectively.

use has a high cost. For this reason, the use of raters with manual stopwatches is usually the most common and feasible option when carrying out measurements of speed–agility tests.[29]

The accuracy with which field-based tests are measured to assess motor fitness using manual stopwatches has been previously demonstrated in adolescents,[13] but there is no evidence in adults aged 18–64 years old.[8] Therefore, it would be of special scientific interest to quantify the error that occurs when these tests are recorded using manual stopwatches in this population. For this, the present study has been performed to analyze the time measurement validity by carrying out two attempts of the 4 × 10-m shuttle run test, measured simultaneously with manual stopwatches (R1 and R2) and automatically with photoelectric cells. In the concordance analysis between the R1 and the gold standard, the systematic error was close to zero, which indicates that there were no differences between these two measurement methods. While, there was a systematic error of 0.2 s between the time recorded by the R2 and the gold standard, although the effect size was <0.1, indicating that there were practically no differences between the standard deviations of the measurements.[28] The amplitude between the LoA was smaller for the R1, which supports and reinforces the aforementioned quantitative results. In addition, there was an excellent degree of agreement between the measurement methods; unlike the R2 where there was a positive proportional bias. Taking into account all mentioned above it would be convenient to use expert rater or train the raters before carrying out the measurements in order to reduce the measurement error and ensure the accuracy of the assessments. Our results concur with those obtained in the analogous study in adolescents,[13] where it was also observed that for the R1 the systematic error was close to zero, whereas for the R2 the error was approximately one tenth of a second. The presence of heteroscedasticity in our study suggests that the degree of agreement between the raters and the photoelectric cells depends on the time required to carry out the test, observing that the less time spent in carrying out the test, the greater the degree of agreement between the measurement methods.

Inter-rater reliability was analyzed through the average difference of the times recorded by the R1 and R2. Although, we observed a systematic error of ~0.18 s, the effect size showed an almost null effect ($d <$ 0.2). However, the distance between the LoA was wide, and there was a presence of a proportional bias, which reinforces the existence of differences between raters, indicating that there was not a good agreement between them. It is important to note that the R2 always recorded longer times, and the degree of agreement between the raters depends on the time spent in carrying out the test, observing that the less time spent in carrying out the test, the greater the agreement between the measurement methods. These results were quite similar to those of the study by Vicente-Rodríguez et al.,[13] although they found a smaller systematic error, just one tenth of a second, and the amplitude between the LoA was also smaller ([0.49, −0.18] vs. [0.37, −0.18]). Since the error is small, but exists, the use of untrained raters should be to rely on scientific judgment, which should decide whether this error is acceptable or not.[30] Perhaps in an epidemiological study with a very large sample size which requires multiple raters to carry it out, and there are not enough trained raters available or the possibility of training them, the use of untrained raters could be accepted, given that the error in the measurements will be greater and, therefore, the reliability would be reduced. Based on the results previously found[13] and the results of our study, this error could not be greater than 0.2 s. The reasonable thing would be a maximum error of 0.1 s, since this error is insignificant, but with 0.2 s the error begins to be significant. According to our study, in 0.1 s there is a displacement of 0.33 m, important for sports performance, but inconsequential in epidemiological studies.

The 4 × 10-m shuttle run test is reliable in youth,[12] however, the test–retest reliability of this test has not been examined in adults.[9] We observed significant differences between T1 and T2. Prior, the presence of a proportional bias might indicate that there was not a good agreement between T1 and T2. However, the systematic error found, although it is significant, was really negligible, only 0.05 s (d ~ 0.03). In addition, the ICC of the Bland–Altman graphical method showed excellent concordance between the measurements, especially when the time required to do the test is low (heteroscedasticity); and all of error measurements presented very low values, close to zero.

Finally, the differences in time recorded between the T1 and the T2 in 81 % of the cases did not exceed 0.3 s, in 59 % of them did not exceed a tenth of a second, and in only 56 % of the cases longer times were recorded in the retest. For all these reasons, we could confirm that there was good test–retest reliability and we were able to rule out the learning effect on the retest measurements.

Regarding the physical activity level groups, we observed that in the active group the time measurement validity and the reliability (between raters and test–retest) improved both for R1 and R2 in terms of systematic error, as well as in the values of mean difference and LoA, and even the CV, which were lower than those observed for the full sample; while it worsens slightly in the non-active group.

Therefore, for the active group, the time measurement validity and the reliability improved regardless of the raters. As we have previously reported, the shorter the time spent on the test, the greater the time measurement validity and the reliability; and in our sample the participants in the active group spent less time than those in the non-active group. Concerning heteroscedasticity, when we divided the sample by physical activity level groups, it disappeared indicating the independence of the time recorded in performing the test with the degree of agreement between the methods and the raters. This could be explained by the fact that by dividing the sample based on specific characteristics, such as the level of physical activity, more homogeneous groups were formed, which makes the variance of the error's constant over time.

4.1. Limitations and strength

This is the first study that analyzes the time measurement validity and reliability (between raters and test–retest) of the 4 × 10-m shuttle run test to assess motor fitness in adults, according to gender, age, and physical activity levels, using a complete set of statistical methods. The lack of control of factors that could affect the test performance, such as genetics or/and experience is a limitation of this study. Another limitation is that almost 50 % of the sample corresponded to young adults (18–34 years), being active all of them; and the mean age of females was older than males (44.56 vs. 29.93 years).

5. Conclusions

The 4 × 10-m shuttle run test appears to be a valid and reliable test in adults when time is measured manually with a stopwatch, reducing the error when time is recorded by a trained rater. It is highly recommended to train the raters in order to minimize the systematic error and ensure the accuracy of the measurements. However, untrained raters could be used in population-based studies as long as the presence of a greater error in the measurements is assumed. The time measurement validity and the reliability of the 4 × 10-m shuttle run test improve when the sample corresponds to physically active adults, regardless of gender and age groups. We consider that the 4 × 10-m shuttle run test is a useful, valid and reliable tool, as well as easy to perform, efficient in time and with low personal cost to assess motor fitness in adults 18–64 years of age.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsams.2023.08.176.

Funding information

This project was supported by the Ministry of Economy, Industry and Competitiveness in the 2017 call for R&D Projects of the State Program for Research, Development and Innovation Targeting the Challenges of the Company; National Plan for Scientific and Technical Research and Innovation 2013–2016 (DEP2017-88043-R); the Spanish Ministry of Education, Culture and Sport (FPU19/02961); and RIG supported by a grant for the Requalification of the Spanish University System (2021–2023) from the Spanish Ministry of Universities (RD 289/2021), funded by the European Union-NextGenerationEU.

Confirmation of ethical compliance

The study was approved by the Committee for Research of Cadiz, Spain.

CRediT authorship contribution statement

JCP, PAA, DCM and JCP contributed to writing the draft manuscript and the statistical analysis. MCC and JCP contributed to the conception and design of the study. JCP, PAA, CCL, JJI, DCM, RIG and MCC contributed to data collection and management. All the authors contributed to the data assessment and approved the final version of the manuscript and agree with the order of presentation of the authors.

Declaration of interest statement

The authors declare that they have no competing interests.

Acknowledgments

The authors thank our participants for the time and effort they volunteered to complete this study.

References

1. Barry VW, Baruth M, Beets MW et al. Fitness vs. fatness on all-cause mortality: a meta-analysis. *Prog Cardiovasc Dis* 2014;56(4):382-390.
2. LaMonte MJ, Barlow CE, Jurca R et al. Cardiorespiratory fitness is inversely associated with the incidence of metabolic syndrome: a prospective study of men and women. *Circulation* 2005;112(4):505-512.
3. Ezzatvar Y, Ramírez-Vélez R, Sáez de Asteasu ML et al. Cardiorespiratory fitness and all-cause mortality in adults diagnosed with cancer systematic review and meta-analysis. *Scand J Med Sci Sports* 2021;31(9):1745-1752.
4. Boots EA, Schultz SA, Oh JM et al. Cardiorespiratory fitness is associated with brain structure, cognition, and mood in a middle-aged cohort at risk for Alzheimer's disease. *Brain Imaging Behav* 2015;9(3):639-649.
5. García-Hermoso A, Cavero-Redondo I, Ramírez-Vélez R et al. Muscular strength as a predictor of all-cause mortality in an apparently healthy population: a systematic review and meta-analysis of data from approximately 2 million men and women. *Arch Phys Med Rehabil* 2018;99(10):2100-2113.e2105.
6. Fraser BJ, Blizzard L, Buscot MJ et al. Muscular strength measured across the life-course and the metabolic syndrome. *Nutr Metab Cardiovasc Dis* 2022;32(5):1131-1137.
7. Marín-Jiménez N, Cruz-León C, Perez-Bey A et al. Predictive validity of motor fitness and flexibility tests in adults and older adults: a systematic review. *J Clin Med* 2022;11(2):328.
8. Castro-Piñero J, Marin-Jimenez N, Fernandez-Santos JR et al. Criterion-related validity of field-based fitness tests in adults: a systematic review. *J Clin Med* 2021;10(16).
9. Cuenca-Garcia M, Marin-Jimenez N, Perez-Bey A et al. Reliability of field-based fitness tests in adults: a systematic review. *Sports Med* 2022;52(8):1961-1979.
10. Ortega FB, Cadenas-Sánchez C, Sánchez-Delgado G et al. Systematic review and proposal of a field-based physical fitness-test battery in preschool children: the PREFIT battery. *Sports Med* 2015;45(4):533-555.
11. Ruiz JR, Castro-Pinero J, Espana-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):518-524.
12. Artero EG, Espana-Romero V, Castro-Pinero J et al. Reliability of field-based fitness tests in youth. *Int J Sports Med* 2011;32(3):159-169.
13. Vicente-Rodríguez G, Rey-López JP, Ruíz JR et al. Interrater reliability and time measurement validity of speed–agility field tests in adolescents. *J Strength Cond Res* 2011;25(7):2059-2063.
14. Rodrigues LP, Luz C, Cordovil R et al. Normative values of the motor competence assessment (MCA) from 3 to 23 years of age. *J Sci Med Sport* 2019;22:1038-1043.
15. Nassif H, Sedeaud A, Abidh E et al. Monitoring fitness levels and detecting implications for health in a French population: an observational study. *BMJ Open* 2012;2(5):e001022.
16. Hals T-MV, Sitler MR, Mattacola CG. Effect of a semi-rigid ankle stabilizer on performance in persons with functional ankle instability. *J Orthop Sports Phys Ther* 2000;30(9):552-556.
17. Nguyen A, Roth N, Ghassemi NH et al. Development and clinical validation of inertial sensor-based gait-clustering methods in Parkinson's disease. *J Neuroeng Rehabil* 2019;16(1):77.
18. Eliakim A, Nemet D, Bar-Sela S et al. Changes in circulating IGF-I and their correlation with self-assessment and fitness among elite athletes. *Int J Sports Med* 2002;23(8):600-603.
19. Nasriah Na NI, Sedek R, Zubairi SI. Relationship between body composition and physical fitness among Royal Malaysia Police personnel in Selangor, Malaysia. *Asian J Clin Nutr* 2017;10(1):25-31.
20. World Health Organization. *Global Recommendations on Physical Activity for Health*, Geneva, World Health Organization, 2010.
21. García-López J, Morante JC, Ogueta-Alday AC et al. El uso de fotocélulas de haz simple y doble para medir la velocidad en carreras®. The use of single- and dual-beam photocells to measure the sprint time®. *RICYDE Revista Internacional de Ciencias del Deporte* 2012;8(30):324-333.
22. Bland JMA, D.G.. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1(8476):307-310.
23. Krouwer JS. Why Bland–Altman plots should useX, not (Y + X)/2 whenX is a reference method. *Stat Med* 2008;27(5):778-780.
24. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19:231-240.
25. Del Corral T, Sánchez ÁG, López-de-Uralde-Villanueva I. Test–retest reliability, minimal detectable change and minimal clinically important differences in modified

shuttle walk test in children and adolescents with cystic fibrosis. *J Cyst Fibros* 2020;19 (3):442-448.
26. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26(4):217-238.
27. Ortega FB, Artero EG, Ruiz JR et al. Reliability of health-related physical fitness tests in European adolescents. The HELENA Study. *Int J Obes (Lond)* 2008;32(Suppl 5):S49-S57.
28. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, London, Routledge, 2013.
29. Hetzler RK, Stickley CD, Lundquist KM et al. Reliability and accuracy of handheld stopwatches compared with electronic timing in measuring sprint performance. *J Strength Cond Res* 2008;22(6):1969-1976.
30. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135-160.