Interrater Reliability and Time Measurement Validity of Speed–Agility Field Tests in Adolescents

Germán Vicente-Rodríguez,[1,2,3] Juan P. Rey-López,[1] Jonathan R. Ruíz,[3,4] David Jiménez-Pavón,[4,5] Patrick Bergman,[3] Donatella Ciarapica,[6] Jose M. Heredia,[7] Denes Molnar,[8] Angel Gutierrez,[4] Luis A. Moreno,[1,9] and Francisco B. Ortega,[3,4] on behalf of the HELENA study group

[1]Growth, Exercise, Nutrition and Development (GENUD) Research Group, University of Zaragoza, Zaragoza, Spain; [2]Faculty of Health and Sport Sciences, Department of Physiotherapy and Nursing, University of Zaragoza, Huesca, Spain; [3]Unit for Preventive Nutrition, Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden; [4]The EFFECTS-262 Research Group, Department of Physiology, School of Medicine, University of Granada, Granada, Spain; [5]Facultad de Ciencias de la Actividad Física y del Deporte (INEF), Universidad politécnica de Madrid, Madrid, Spain; [6]Istituto Nazionale di Ricerca per gli Alimenti e la Nutrizione Human Nutrition Unit INRAN, Roma, Italy; [7]Departamento de Educación Física y Deportiva, Universidad de Granada, Granada, Spain; [8]University of Pécs, Department of Paediatrics, Pécs-JózsefA; and [9]School of Health Sciences, Department of Physiotherapy and Nursing, University of Zaragoza, Zaragoza, Spain

Abstract

Vicente-Rodríguez, G, Rey-López, JP, Ruíz, JR, Jiménez-Pavón, D, Bergman, P, Ciarapica, D, Heredia, JM, Molnar, D, Gutierrez, A, Moreno, LA, and Ortega, FB. Interrater reliability and time measurement validity of speed–agility field tests in adolescents. *J Strength Cond Res* 25(X): 000–000, 2011—The aim of this study was to examine the interrater reliability (trained vs. untrained raters) and criterion-related validity (manual vs. automatic timing) of the 4 × 10-m shuttle run and 30-m running speed tests (times measured). The study comprised 85 adolescents (38 girls) aged 13.0–16.9 years from the Healthy Lifestyle in Europe by Nutrition in Adolescence study. The time required to complete the 4 × 10-m shuttle run and 30-m running tests was simultaneously measured (a) manually with a stopwatch by both trained and untrained raters (for interrater reliability analysis), and (b) by using photoelectric cells (for validity analysis). Systematic error, random error, and heteroscedasticity were studied with repeated-measured analysis of variance and Bland–Altman plots. The systematic error for untrained vs. trained raters and the untrained raters vs. photoelectric cells were in all cases ~0.1 seconds ($p < 0.01$), that is, untrained raters recorded higher times. No systematic error was found between trained raters and photoelectric cells ($p > 0.05$). No heteroscedasticity was shown in any case ($p > 0.05$). The findings indicate that manual measurements by a trained rater, using a stopwatch, seem to be a valid method to assess speed and agility fitness testing in adolescents. Researchers must be trained to minimize the measurement error.

Key Words health, fitness, velocity, exercise, methodology

Introduction

Physical fitness is a marker of health in children and in adolescents (10,12). Muscular fitness and speed/agility are important physical components related to youth health status; however, they have been under studied in the literature, in comparison to cardiorespiratory function (10).

Recent systematic reviews have identified a need for thorough validity and reliability studies on fitness testing young populations, particularly in speed/agility tests (3,12). Field-based fitness tests such as the 4 × 10-m shuttle run and 30-m running speed tests are used worldwide in population-based studies, sport centers, and schools; however, whether these tests are valid and reliable is not known.

Field-based speed and agility tests are usually evaluated with manual stopwatches, yet no information is available about the resultant error when compared with a gold standard for time-measuring in fitness testing, such as photoelectric cells, or about the error between trained vs. untrained raters. We hypothesized that the time measurement in speed and agility tests (i.e., 4 × 10-m shuttle run and 30-m running speed tests) can be accurately measured using a manual stopwatch if the rater is

Address correspondence to Dr. Germán Vicente-Rodríguez, gervicen@unizar.es.

0(0)/1–5

Journal of Strength and Conditioning Research

properly trained. A systematic error can be however present if the time measurement is carried out by an untrained rater.

This study aimed to contribute to the current knowledge on pediatric fitness testing by analyzing the interrater reliability (trained vs. untrained rater) and criterion-related validity (manual vs. automatic timing) of the times measured in the 4 × 10-m shuttle run and 30-m running speed tests in the adolescent population. The results from these analyses will help sport scientists and practitioners to understand how large the error is when using stopwatches in speed–agility testing and to what extent the raters need to be trained to reduce this error.

Methods

Experimental Approach to the Problem

For a test to be considered ‘good,’ it should measure what it purports to measure (i.e., validity) with consistency (i.e., reliability). To our knowledge, there are no studies examining the validity and reliability of the 4 × 10-m shuttle run and 30-m running tests in adolescents despite the use of these measurements in many field-based fitness test batteries currently used worldwide to assess speed and agility (3). This study assessed the interrater reliability for the 4 × 10-m shuttle run and 30-m running speed tests by comparing the time measured by a trained and untrained rater. Criterion-related validity was studied comparing the time measured by the raters (both trained and untrained) with the time recorded by the photoelectric cells (the gold standard).

Subjects

The Healthy Lifestyle in Europe by Nutrition in Adolescence (HELENA) study is a European Union-funded project (7), which includes a cross-sectional multicenter study. From the HELENA cross-sectional study sample recruited in Zaragoza (Aragón, Spain), a total of 85 (38 girls) adolescents aged 13–16.9 years voluntarily participated in this study. The study was approved by the Research Ethics Committee of the Government of Aragón (Spain), and written informed consent was gained from both parents and adolescents.

Procedures

Physical Fitness Testing. **4 × 10-m shuttle run.** This test was used to assess speed, agility and coordination (8). Two parallel lines were drawn on the floor 10 m apart. Subjects ran back and forth as fast as possible crossing each line with both feet every time. This was performed twice, covering a total distance of 40 m (4 × 10 m). Every time the adolescent crossed any of the lines, he or she picked up (the first time) or exchanged (second and third time) a sponge, which was previously placed behind the lines. The stopwatch (manual or photoelectric cells) was stopped when the adolescent crossed the end line with one foot. The time required to complete the test was recorded to the nearest tenth of a second. A slip-proof floor, 4 cones, a stopwatch and 3 sponges were used to perform the test.

30-m running speed test. This test was used to assess speed (13). We measured the time required to cover 30 m on a straight track. At the starting line, participants stood in a stationary and comfortable position with their feet behind the starting line (where the first photoelectric barrier was placed), with no rocking movements. Runners were instructed to start on the whistle sound.

All the tests were performed twice with at least 1 minute of rest between attempts. The participants received thorough instructions after which they were also allowed to practice the tests. Subjects received verbal encouragement during the tests, and the best score was retained and used in analysis.

Experiment Design

Timekeeper Selection. There were 2 timekeepers: a trained and an untrained rater. The trained rater (R1) was the timekeeper who had previously participated in the HELENA training period (9) and pilot study. The untrained rater (R2) did not attend the training period and did not participate in the HELENA pilot study.

Interrater Reliability. The time required to complete each test was measured simultaneously with a stopwatch by R1 and R2. Raters started the stopwatch at the whistle sound (given by a third researcher) and stopped the stopwatch as the runner crossed through the finishing line.

Validity of the Manual Timer Measurement. The 4 × 10-m shuttle run and 30-m running speed tests were also automatically measured with photoelectric cells (Byomedics, Barcelona). For sport and fitness tests, photoelectric cells are used as a reference measurement to accurately assess the time required to cross 2 separate lines. The photoelectric cell timer was automatically activated when the runner crossed the first cell and stopped when the subject crossed the last cell. For the 4 × 10-m shuttle run the photoelectric cells were placed at the start/finishing line (it was the same line), where the raters were also placed. For the 30-m running speed test, photoelectric cells were placed at the start and finishing lines. Raters were placed 15 m from the start line.

Statistical Analyses

All the study variables were normally distributed. The analyses were performed using SPSS v.15 software for windows and the significance level was set at $p \leq 0.05$ for all the analysis. The degree of agreement between times measured by the trained vs. untrained rater (interrater agreement) were graphically examined by plotting the difference between raters against the gold standard (photoelectric cells), according to Bland–Atlman method (1). The agreement between methods (manual stopwatch and photoelectric cells) was examined graphically by plotting the difference between each rater and photoelectric cells. Differences were plotted against the gold standard instead of the mean value because the gold standard was expected to be closer to the "true value" than the mean (6). Additionally,

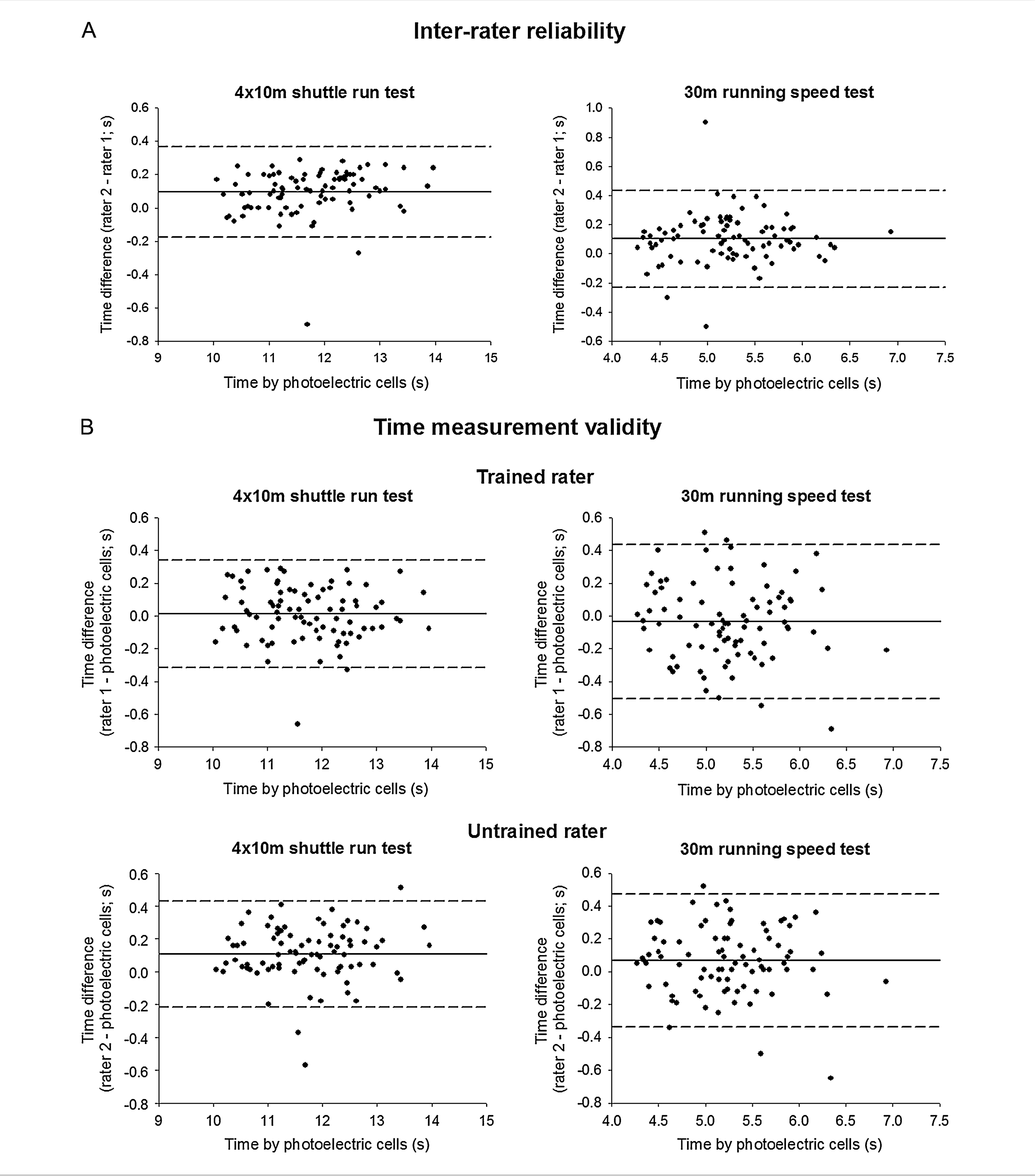

Figure 1. Bland–Altman plots for A) interrater reliability showing the mean difference between untrained rater (rater 2) − trained rater (rater 1) vs. time measured by photoelectric cells for 4 × 10-m shuttle run and 30-m running speed test. B) Time measurement validity in trained and untrained raters for the 4 × 10-m shuttle run and 30-m running speed 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*SD*s of the differences), respectively.

the presence of systematic error was analyzed by a repeated-measured analysis of variance (ANOVA). The pairs of factors included for the repeated-measured ANOVA were R1 vs. R2, R1 vs. photoelectric cells, and R2 vs. photoelectric cells. The limits of agreement were computed as systematic error (mean difference) ± 1.96*SD*s (of the mean differences).

The presence of heteroscedasticity was additionally studied by 1-way ANOVA, setting the absolute difference (negative

Table 1. Interrater reliability (untrained vs. trained raters) and time measurement validity (raters vs. photoelectric cells) statistics for the 4 × 10-m shuttle run and 30-m running speed tests in adolescents.

	Systematic error*	95% Limits of agreement†	
Interrater reliability (untrained–trained raters)			
4 × 10-m Shuttle run (s)	0.09‡	−0.18	0.37
30-m Running speed (s)	0.10‡	−0.23	0.43
Time measurement validity for the trained raters (trained raters–cells)			
4 × 10-m Shuttle run (s)	0.01	−0.31	0.34
30-m Running speed (s)	−0.04	−0.51	0.43
Time measurement validity for the untrained raters (untrained raters–cells)			
4 × 10-m Shuttle run (s)	0.11‡	−0.22	0.43
30-m Running speed (s)	0.07§	−0.34	0.47

*Systematic error, mean difference.
†95% Limits of agreement; mean difference ± 1.96 *SD*s of the differences.
‡$p < 0.001$.
§$p < 0.01$, analyzed by ANOVA of the interrater or intermethods differences.

values were multiplied by -1) as dependent variable and quartiles of magnitude (in this case photoelectric cells) as factor (9).

Results

Interrater Reliability

The interrater reliability patterns between timekeepers for each test are graphically presented in Figure 1A, which shows the untrained rater always recorded higher times than the trained rater. No heteroscedasticity (change in the variance by increasing or decreasing the magnitude of the measurement) was observed (data not shown). The ANOVA of the interrater differences revealed systematic error for both fitness tests ($p < 0.01$, Table 1).

Validity of Manual Time Measurement in 4 × 10-m Shuttle Run and 30-m Running Speed Tests

Validity for the 4 × 10-m shuttle run and 30-m running speed tests when simultaneously measured by the raters and photoelectric cells is graphically shown in Figure 1B. No heteroscedasticity was found. We observed a systematic error of ~0.1 seconds for the untrained raters ($p < 0.001$ and <0.01, for 4 × 10 m and 30-m test, respectively), whereas it was nearly zero for the trained raters (Table 1).

Discussion

The major finding of this study is that both the 4 × 10-m shuttle run and 30-m running speed tests provide valid results when assessed by a trained timekeeper. The use of these tests is of interest not only for research but also in educational and sport centers. Although the ideal option would be to use photoelectric cells, the most commonly used measurement tool is a stopwatch because it is much cheaper and feasible in certain settings, such as schools and population-based studies (5). The precision and reliability of sprint performance was scarcely studied in adults (4,5,11) before the advent of digital quartz stopwatches and commercially available photoelectric cells (5). To our knowledge, the precision of timing speed and agility tests with a manual stopwatch, and the interrater reliability has not been previously studied in adolescents (3). Therefore, from a scientific point of view it is of interest to know the amount of error that occurs when these field tests are performed with stopwatches. The first step is to examine whether this timing method is valid. Thus, we compared time required to perform 2 running (i.e., the 4 × 10-m shuttle run and 30-m test) tests when simultaneously measured manually with a stopwatch and automatically with photoelectric cells. Results showed there was no systematic error in trained raters, whereas the times were significantly overestimated when measured by untrained raters. The lack of heteroscedasticity suggests the degree of agreement between the trained rater and photoelectric cells is independent of the time taken to perform the tests.

We also studied the mean difference measured by trained and untrained raters. The results showed that there was systematic error. The decision about what is 'acceptable' should be based on scientific judgment, because measurements that agree well enough for one purpose may not agree well enough for another (2). The findings observed in this study indicate differences in mean time scored between trained and untrained raters were statistically significant, and the differences in mean time scored between the untrained rater and the photoelectric cells. However, these differences were roughly one-tenth of a second for the 4 × 10-m shuttle run and the 30-m running speed tests (which last around 11 and 5 seconds, respectively), with the untrained rater always recording longer times. Greater reliability (smaller systematic error) was observed between the trained rater and the photoelectric cells.

In light of these findings, we believe such a small systematic error could be assumed for a mean value in large cohort studies at that population level, yet not at individual levels (5). In a prior study carried out with trained raters, we observed a good test–retest agreement for a set of physical fitness tests, and neither learning nor fatigue effect of the subjects was noted for any of the tests studied (9). These findings together with those observed in this study suggest that to achieve high scientific standards at a population level or high accuracy at individual level, researchers must be properly trained to ensure a valid and reliable measurement of speed/agility.

In conclusion, time measured manually by a stopwatch seems to be a valid and reliable method for measuring speed and agility in large cohort studies involving adolescents. This study also suggests that the training of the researchers is highly recommended to minimize systematic error and to ensure accurate measurements.

Practical Applications

Researchers, coaches, and physical education teachers can acceptably measure the time required to performed speed and agility tests using a stopwatch, even when they are not trained for it. However, if a more precise measurement is needed, that can be attained through training with a stopwatch in speed/agility field-based tests.

Acknowledgments

We gratefully acknowledge the collaboration of all of our subjects, their parents and teachers. The authors wish to thank Prof. Olle Carlsson for his assistance with the statistical analysis of the data. The HELENA study received financial support from the European Community Sixth Framework Program (Contract FOOD-CT-2005-007034). The Assessing Levels of Physical Activity and Fitness at Population Level (ALPHA) study received funding from the European Union, within the framework of the Public Health Program (Ref: 2006120). G.V.R. was supported by a grant from the Dirección General de Aragón (DGA; CONSI+D) and Caja de Ahorros de la Inmaculada (CAI; CM 11/07). This study is also being supported by grants from the Spanish Ministry of Education (EX-2007-1124, EX-2008-0641), Swedish Council for Working Life and Social Research, and the ALPHA study, a European Union-funded study, in the framework of the Public Health Program (Ref: 2006120). D.J. is supported by a grant from the Spanish Ministry of Education (AP2005-3827). J.P.R. is supported by a grant from Fundación Cuenca Villoro. The content of this article reflects only the authors' views, and the European Community is not liable for any use that may be made of the information contained therein.

References

1. Bland, JM and Altman, DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1: 307–310, 1986.
2. Bland, JM and Altman, DG. Measuring agreement in method comparison studies. *Stat Meth Med Res* 8: 135–160, 1999.
3. Castro-Pinero, J, Artero, EG, Espana-Romero, V, Ortega, FB, Sjostrom, M, Suni, J, and Ruiz, JR. Criterion-related validity of field-based fitness tests in youth: A systematic review. *Br J Sports Med* 44: 934–943, 2010..
4. Fry, AC and Kraemer, WJ. Physical performance characteristics of American collegiate football players. *J Appl Sports Sci* 5: 126–138, 1991.
5. Hetzler, RK, Stickley, CD, Lundquist, KM, and Kimura, IF. Reliability and accuracy of handheld stopwatches compared with electronic timing in measuring sprint performance. *J Strength Cond Res* 22: 1969–1976, 2008.
6. Krouwer, JS. Why Bland–Altman plots should use X, not $(Y+X)/2$ when X is a reference method. *Stat Med* 27: 778–780, 2008.
7. Moreno, LA, De Henauw, S, Gonzalez-Gross, M, Kersting, M, Molnar, D, Gottrand, F, Barrios, L, Sjostrom, M, Manios, Y, Gilbert, CC, Leclercq, C, Widhalm, K, Kafatos, A, and Marcos, A. Design and implementation of the healthy lifestyle in Europe by nutrition in adolescence cross-sectional study. *Int J Obes (Lond)* 32(Suppl 5): S4–S11, 2008.
8. Ortega, FB, Artero, EG, Ruiz, JR, Espana-Romero, V, Jimenez-Pavon, D, Vicente-Rodriguez, G, Moreno, LA, Manios, Y, Beghin, L, Ottevaere, C, Ciarapica, D, Sarri, K, Dietrich, S, Blair, SN, Kersting, M, Molnar, D, Gonzalez-Gross, M, Gutierrez, A, Sjostrom, M, and Castillo, MJ. Physical fitness levels among European adolescents: The HELENA study. *Br J Sports Med* 2011.
9. Ortega, FB, Artero, EG, Ruiz, JR, Vicente-Rodriguez, G, Bergman, P, Hagstromer, M, Ottevaere, C, Nagy, E, Konsta, O, Rey-Lopez, JP, Polito, A, Dietrich, S, Plada, M, Beghin, L, Manios, Y, Sjostrom, M, and Castillo, MJ. Reliability of health-related physical fitness tests in European adolescents. The HELENA Study. *Int J Obes (Lond).* 32(Suppl 5): S49–S57, 2008.
10. Ortega, FB, Ruiz, JR, Castillo, MJ, and Sjostrom, M. Physical fitness in childhood and adolescence: A powerful marker of health. *Int J Obes (Lond)* 32: 1–11, 2008.
11. Reid, DD and Sandland, RL. New lamps for old? *J R Stat Soc Ser C Appl Stat* 32: 86–87, 1983.
12. Ruiz, JR, Castro-Pinero, J, Artero, EG, Ortega, FB, Sjostrom, M, Suni, J, and Castillo, MJ. Predictive validity of health-related fitness in youth: A systematic review. *Br J Sports Med* 43: 909–923, 2009.
13. Vicente-Rodriguez, G, Jimenez-Ramirez, J, Ara, I, Serrano-Sanchez, JA, Dorado, C, and Calbet, JA. Enhanced bone mass and physical fitness in prepubescent footballers. *Bone* 33: 853–859, 2003.