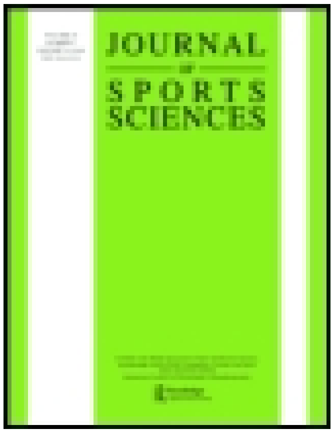


Journal of Sports Sciences
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjsp20

Are the Yo-Yo intermittent recovery test levels 1 and 2 both useful? Reliability, responsiveness and interchangeability in young soccer players

Maurizio Fanchini[ab], Carlo Castagna[c], Aaron J. Coutts[d], Federico Schena[a], Alan McCall[f] & Franco M. Impellizzeri[e]

[a] Faculty of Exercise and Sport Science, University of Verona, Verona, Italy

[b] FC Internazionale, Milan, Italy

[c] Football Training and Biomechanics Laboratory, Italian Football Association, Florence, Italy

[d] Faculty of Health, University of Technology Sydney (UTS), Sydney, Australia

[e] Research & Development, Schulthess Clinic, Zurich, Switzerland

[f] Universite de Lille2, Lille, France

Published online: 21 Oct 2014.

To cite this article: Maurizio Fanchini, Carlo Castagna, Aaron J. Coutts, Federico Schena, Alan McCall & Franco M. Impellizzeri (2014): Are the Yo-Yo intermittent recovery test levels 1 and 2 both useful? Reliability, responsiveness and interchangeability in young soccer players, Journal of Sports Sciences, DOI: 10.1080/02640414.2014.969295

To link to this article: http://dx.doi.org/10.1080/02640414.2014.969295

Journal of Sports Sciences, 2014
http://dx.doi.org/10.1080/02640414.2014.969295

Routledge
Taylor & Francis Group

Are the Yo-Yo intermittent recovery test levels 1 and 2 both useful? Reliability, responsiveness and interchangeability in young soccer players

MAURIZIO FANCHINI[1,2], CARLO CASTAGNA[3], AARON J. COUTTS[4], FEDERICO SCHENA[1], ALAN McCALL[6] & FRANCO M. IMPELLIZZERI[5]

[1]*Faculty of Exercise and Sport Science, University of Verona, Verona, Italy,* [2]*FC Internazionale, Milan, Italy,* [3]*Football Training and Biomechanics Laboratory, Italian Football Association, Florence, Italy,* [4]*Faculty of Health, University of Technology Sydney (UTS), Sydney, Australia,* [5]*Research & Development, Schulthess Clinic, Zurich, Switzerland and* [6]*Universite de Lille2, Lille, France*

(*Accepted 19 September 2014*)

Abstract

The aim of this study was to compare the reliability, internal responsiveness and interchangeability of the Yo-Yo intermittent recovery test level 1 (YY1), level 2 (YY2) and submaximal YY1 (YY1-sub). Twenty-four young soccer players (age 17 ± 1 years; height 177 ± 7 cm; body mass 68 ± 6 kg) completed each test five times within pre- and in-season; distances covered and heart rates (HRs) were measured. Reliability was expressed as typical error of measurement (TEM) and intraclass correlation coefficient (ICC). Internal responsiveness was determined as effect size (ES) and signal-to-noise ratio (ES_{TEM}). Interchangeability was determined with correlation between training-induced changes. The TEM and ICC for distances in the YY1 and YY2 and for HR in YY1-sub were 7.3% and 0.78, 7.1% and 0.93 and 2.2% and 0.78, respectively. The ESs and ES_{TEMs} were 0.9 and 1.9 for YY1, 0.4 and 1.2 for YY2 and −0.3 and −0.3 for YY1-sub. Correlations between YY1 vs. YY2 and YY1-sub were 0.56 to 0.84 and −0.36 to −0.81, respectively. Correlations between change scores in YY1 vs. YY2 were 0.29 and −0.21 vs. YY1-sub. Peak HR was higher in YY1 vs. YY2. The YY1 and YY2 showed similar reliability; however, they were not interchangeable. The YY1 was more responsive to training compared to YY2 and YY1-sub.

Keywords: *team sport, soccer test, reproducibility, sensitivity to changes, convergent validity*

Introduction

The Yo-Yo intermittent recovery test (YYIRT) is a valid field test and widely used in soccer. Studies have shown that it is correlated with the high-intensity activity performed during a match and differentiates between competitive levels, playing positions and changes throughout the season (Bangsbo, Iaia, & Krustrup, 2008; Mohr & Krustrup, 2014; Mujika, Santisteban, Impellizzeri, & Castagna, 2009). The widespread use of the YYIRT is likely due also to its relative simplicity, low cost and capacity to test several players at the same time.

Two versions of the YYIRT (levels 1 and 2, YY1 and YY2, respectively) have been proposed (Krustrup et al., 2003, 2006). Several studies have suggested that the YY1 performance is more dependent on aerobic system compared to the YY2 (Ingebrigtsen et al., 2012; Rampinini et al., 2010). However, the YY2 induces a higher contribution from the anaerobic system compared with YY1 (Bangsbo et al., 2008; Rampinini et al., 2010). Given that these are maximal tests, which are fatiguing, the submaximal version of the YY1 (YY1-sub) was developed as a practical alternative that would allow frequent fitness assessments, which could be easily incorporated into training or rehabilitation plans (Bangsbo et al., 2008). The outcome measure of this version is the HR taken at the 6th min (HR6th) of the YY1 (Krustrup et al., 2003). The HR6th in the YY1-sub has shown a moderate correlation with high-intensity running covered during a match (r = −0.48) (Bangsbo et al., 2008) and decreases throughout the season showing an improvement in the test values (Mohr & Krustrup, 2014).

Although correlations between the different versions of the YYIRTs have been investigated (Ingebrigtsen et al., 2012, 2014; Karakoç, Akalan, Alemdaroğlu, & Arslan, 2012; Mohr & Krustrup,

Correspondence: Maurizio Fanchini, Faculty of Exercise and Sport Science, University of Verona, Verona, Italy. E-mail: maurizio.fanchini@gmail.com

2014; Rampinini et al., 2010), no studies have examined whether they measure different physical capacities or whether they are interchangeable. For example, whilst the YY1-sub has been proposed as a substitute of YY1 (Bangsbo et al., 2008) to the best of our knowledge, no studies have examined whether a change in the HR6th reflects a change in the maximal version (i.e. external responsiveness) (Impellizzeri & Marcora, 2009). Similarly, whilst the YY1 and YY2 both measure the ability to perform high-intensity intermittent exercise, they have been suggested to assess different physiological responses (e.g. different aerobic and anaerobic contribution). It is not currently known whether these tests provide similar information, which may leave one of the tests redundant. Therefore, the correlation between changes in the two tests should be examined (with high shared variance between changes indicating redundancy and interchangeability). From a practical perspective, this information is very important, as it would help in understanding whether or not it is necessary to use both tests.

The selection of a test should be based on its measurement properties and purpose for assessment (e.g. monitoring training or selection and comparisons). In order to examine the appropriateness of a test to measure changes over time (due to training interventions), another important measurement attribute is the sensitivity to changes or internal responsiveness (Impellizzeri & Marcora, 2009). Responsiveness concerns the ability of a test to detect changes, it is also linked to reliability, and both are essential components in the validation process of every measurement tool (Aaronson et al., 2002). The acceptability of the level of reliability cannot be determined using benchmarks but rather must be based on the changes (or the differences) we would like to track. The test should be able to distinguish important (worthwhile) changes from measurement error. Using this approach, the smallest worthwhile change (SWC; i.e. the smallest changes that can be considered important) should be higher than the minimal detectable change (MDC; i.e. the minimal changes considered real with an acceptable probability level) so that small but important changes at an individual level can be detected.

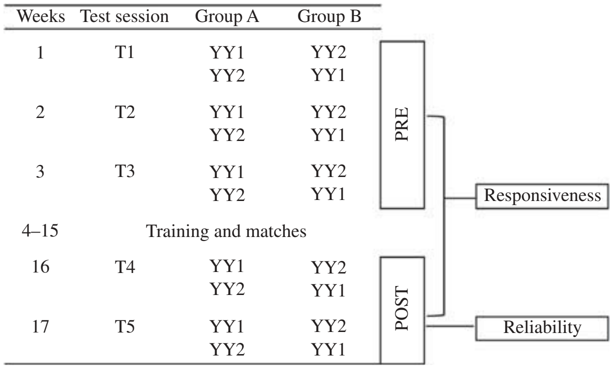
In addition, interpreting the reliability as the measurement noise and the training-induced changes (or other intervention) as the signal, the responsiveness at group level can be calculated as signal-to-noise ratio (ES_{TEM}) (Amann, Hopkins, & Marcora, 2008; Norman, Wyrwich, & Patrick, 2007). Furthermore, two types of reliability can be identified and reported: absolute (also called agreement) (de Vet, Terwee, Knol, & Bouter, 2006), which is important for longitudinal assessment, and relative reliability (or simply reliability), which is useful for discriminant ability between individuals (Impellizzeri & Marcora, 2009). To the best of our knowledge, these aforementioned measurement attributes have not been purposely investigated at least using quantitative and not qualitative methods (Impellizzeri & Marcora, 2009).

The purpose of this investigation was to compare the different versions of the YYIRTs by examining: absolute and relative reliability, SWCs, MDCs, internal responsiveness and their interchangeability. In addition, the external responsiveness for the YY1-sub was examined. We hypothesised that the YYIRTs would be reliable and responsive to soccer training but not interchangeable.

Methods

Participants and study design

Twenty-four junior players (age 17 ± 1 years; height 177 ± 7 cm; body mass 68 ± 6 kg) from a professional fourth division Italian soccer team participated in the study. During the first 2 days of pre-season (PRE), players carried out a general introductory program based on low-intensity running, stretching and technical exercises. Players were then randomly divided into two groups and performed both the YY1 and YY2 once per week for 3 weeks. The tests were performed following randomisation of the starting order in alternate sequence and with at least 48 h between tests resulting in three sessions for each level of YYIRTs in PRE (T1–T2–T3). After 11 weeks (i.e. 3 weeks of training and 8 weeks of training and matches, POST), the YY1 and YY2 were repeated again twice (T4 and T5) (Figure 1). Before all tests, players completed an identical standardised light warm-up consisting of jogging, dynamic and static stretching of the lower limbs and shuttle runs. During the testing period (T1–T2–T3 and T4 and T5), the other sessions included low-intensity exercises as the high-intensity exercises



Weeks	Test session	Group A	Group B
1	T1	YY1 YY2	YY2 YY1
2	T2	YY1 YY2	YY2 YY1
3	T3	YY1 YY2	YY2 YY1
4–15		Training and matches	
16	T4	YY1 YY2	YY2 YY1
17	T5	YY1 YY2	YY2 YY1

Figure 1. Schematic representation of the study and test sequences.

were done performing the tests. Training between PRE and POST consisted of interval training (Impellizzeri et al., 2006), sprint training (Ferrari Bravo et al., 2008), small-sided games (Impellizzeri et al., 2006; Rampinini et al., 2007), as well as tactical and technical exercises. During the season (as between PRE and POST), all players trained 4 days per week (average duration 80–90 min) and played one official match (90 min) during the weekend.

Before participating, all players and their parents provided a written informed consent. The study was approved by the Ethics Committee of the University of Verona.

Yo-Yo intermittent recovery tests

The players completed the YY1 and YY2 as described by Krustrup et al. (2003, 2006). Briefly, the players completed 20-m shuttle runs at increasing velocities with 10 s of active recovery between runs until exhaustion. The YY1 started at 10 km h^{-1} and the YY2 at 13 km h^{-1}. Both tests were terminated when the player failed to reach the line on time on two consecutive occasions. According to Krustrup et al. (2003, 2006), the tests were performed using an audio signal. HR data were collected using a long-range telemetry system (Suunto t6, Team Pack Pro, Suunto Team Pod, Suunto Oy). The maximal distances reached in the YY1 and YY2 and the HR6th in the YY1 (YY1-sub) were used as the outcome measures.

Statistical analysis

All data were presented as mean ± standard deviation (mean ± *s*). The assumption of normality was verified using the Shapiro–Wilk test. Effects and differences (percentage changes T1 vs. T2, T2 vs. T3, T4 vs. T5 and best performance PRE vs. POST) were presented with their corresponding 90% confidence intervals (CI) and calculated after log transformation of the data to reduce bias due to nonuniformity error. Changes between the tests were examined with a repeated measures ANOVA. When a significant *F*-value was found, the Bonferroni *post hoc* test was applied. Effect size (ES) (partial eta-squared, η^2) was also calculated and values of 0.01, 0.06 and above 0.15 were interpreted as small, medium and large, respectively. Absolute reliability was assessed using the typical error of measurement (TEM) expressed as percentage coefficient of variation (CV) and 90% CIs (Hopkins, 2000). Relative reliability was assessed using the intraclass correlation coefficients [ICC (2.1), a two-way random effects model with single measure] and 90% CIs (Weir, 2005). The MDC (expressed as percentage) was considered as the minimal changes in the tests that can be interpreted as real (outside the TEM) and was calculated using the formula: $1.96 \cdot \sqrt{2} \cdot \text{TEM}$ (de Vet et al., 2006), where 1.96 derives from the 95% CI of no change and √2 because of the difference of two variances (Beckerman et al., 2001). The SWC was calculated using a distribution-based method, that is, as a proportion of the effect size that represents the magnitude of improvement in a variable as a function of the between-participants standard deviation of the investigated population (i.e. 0.2 times the between-participant *s*) (de Vet et al., 2006). Internal responsiveness was measured using the best score of PRE and POST tests according to two methods: (1) in relation to the baseline inter-participant variability (Cohen's ES) calculated as the mean difference between POST and PRE test scores, divided by the *s* of baseline scores (Husted, Cook, Farewell, & Gladman, 2000); (2) in relation to the TEM (ES$_{\text{TEM}}$) calculated as the mean difference between the POST and PRE test scores, divided by the TEM (Amann et al., 2008; Norman et al., 2007). Corresponding 90% CIs of the effect sizes (ES and ES$_{\text{TEM}}$) were calculated using the spreadsheet provided by Hopkins (2007) (http://newstats.org/xcl.xls). In addition, we calculated the probability of substantial changes between best score in PRE and POST (i.e. larger than the SWC) (Batterham & Hopkins, 2006; Impellizzeri et al., 2008). Thresholds for assigning qualitative terms to the changes were as follows: <1%, almost certainly not; <5%, very unlikely; <25%, unlikely or probably not; <75%, possibly may not; <95%, likely, probable; <99%, very likely, almost certain (Impellizzeri et al., 2008; Liow & Hopkins, 2003). Pearson product moment of correlation (90% CI) was calculated to examine relationships between tests and to examine the correlation between change scores (YY1 vs. YY2, YY1 vs. YY1-sub and YY2 vs. YY1-sub). As for internal responsiveness, the best performance reached in PRE and POST was used to assess correlation between change scores (interchangeability). The magnitude of the correlations was determined according to Hopkins as follows (http://www.sportsci.org/resource/stats/2002): r < 0.1, trivial; 0.1–0.3, small; 0.3–0.5, moderate; 0.5–0.7, large; 0.7–0.9, very large; >0.9, nearly perfect; and 1 perfect. In addition, the HR reached in every test session was measured in order to compare if the two tests are interchangeable in the detection of the peak HR. Statistical significance was set at the conventional level: $P \le 0.05$. For statistical analysis, the spreadsheet provided by Hopkins (www.sportsci.org, http://www.sportsci.org/resource/stats/relycalc.html) and SPSS software (Version 13.0, SPSS Inc., Chicago, IL, USA) were used.

Results

Absolute and relative reliability

The outcomes of the tests, the ANOVA and *post hoc* results are presented in Table I. The distance covered and the peak HR were significantly different between the five sessions in all tests (all $P \leq 0.05$).

TEM and ICC are presented in Table II. Despite the absence of statistically significant changes between T1 and T2 tests, the scores showed a trend to increase (Table I), which could be due to a learning or training effect. Therefore, only the POST tests (T4 and T5) were used in the reliability assessment (Figure 1), although the reliability calculated using the PRE tests was similar (data not shown). The maximal tests showed similar absolute reliability. The YY2 showed higher relative reliability compared to YY1 and YY1-sub. The YY1-sub showed higher reproducibility compared to YY1 and YY2.

SWC and MDC

The SWC and MDC values expressed as percentage are presented in Table II. The overall SWC values considered as the minimal change that players should reach in future tests were 66.9 m and 33.2 m for YY1 and YY2 and 1.6 beats · min^{-1} for YY1-sub.

Changes after training and internal responsiveness

The best performances in PRE and POST and changes between tests are presented in Table II. Percentage changes after the training period, chances of substantial changes and qualitative descriptors are presented in Figure 2. The HR between the T1 and T5 tests should be considered in order to study changes after training in YY1-sub. However, due to a tendency to increase in the values of the YY1, we could not exclude the presence of a learning or training effect between T1 and T2. This may have influenced the running activity and consequently HR6th. For this reason, the training-induced changes and

Table I. Distance and peak heart rate in the YY1 and YY2. Heart rate at 6th min in the YY1-sub.

	Test 1 Mean ± s	Test 2 Mean ± s	Test 3 Mean ± s	Test 4 Mean ± s	Test 5 Mean ± s	ANOVA P-level	Partial η^2
YY1							
Distance (m)	1668 ± 256	1795 ± 322	1856 ± 255[a]	2082 ± 312[a,b,c]	2130 ± 298[a,b,c]	<0.0001	0.6
Peak heart rate (beats · min^{-1})	198 ± 9	196 ± 8	195 ± 7[d]	196 ± 8	195 ± 8	0.05	0.1
YY1-sub (HR6th)	186 ± 7	180 ± 8[a]	176 ± 7[a,b]	180 ± 8[a]	177 ± 8[a]	<0.0001	0.5
YY2							
Distance (m)	645 ± 144	660 ± 163	696 ± 125	773 ± 184[a,b]	747 ± 181[a]	0.002	0.3
Peak heart rate (beats · min^{-1})	194 ± 7	189 ± 7[d]	189 ± 7[d]	192 ± 7	193 ± 7[e,f]	<0.0001	0.3

Note: Mean ± s values of distances, peak heart rates in YY1, YY1-sub and YY2 in all test sessions. Significant differences were a: vs. Test 1, b: vs. Test 2, c: vs. Test 3, (all $P \leq 0.004$), d: vs. Test 1 ($P < 0.05$), e: vs. Test 2 ($P = 0.02$), f: vs. Test 3 ($P = 0.05$).

Table II. Comparison of the YY1, YY2 and YY1-sub. PRE and POST best distances performed in YY1 and YY2 and heart rates at the 6th min for the YY1-sub. Percentage training-induced changes (distances and HRs in maximal and YY1-sub test, respectively) and reliability (absolute and relative), internal responsiveness, smallest worthwhile change and minimal detectable change for all the tests.

STATISTICS	YY1	YY2	YY1-sub
Best performance PRE (m)	1911 ± 268	718 ± 141	
Best performance POST (m)	2188 ± 298	789 ± 184	
Peak heart rate PRE (beat·min^{-1})	197 ± 7	191 ± 7	
Peak heart rate POST (beat·min^{-1})	196 ± 7	194 ± 7	
Heart rate 6th PRE (beat·min^{-1})			180 ± 8
Heart rate 6th POST (beat·min^{-1})			177 ± 8
Percentage changes after training (90% CI)	14.5 (10.3 to 19.0)	8.8 (3.4 to 14.4)	−1.3 (−2.4 to −0.2)
Probability of substantiality of the changes (%)	99.9/0.1/0	89.6/10.4/0	0/27/75
Reliability			
Absolute (TEM as CV%, 90% CI)	7.3 (5.8 to 9.8)	7.1 (5.7 to 9.5)	2.2 (1.7 to 2.9)
Relative (ICC) (2.1) (90% CI)	0.78 (0.61 to 0.89)	0.93 (0.86 to 0.96)	0.78 (0.60 to 0.88)
Responsiveness (effect size)			
ES (90% CI)	0.9 (0.66 to 1.18)	0.4 (0.17 to 0.69)	−0.29 (−0.54 to −0.04)
ES_{TEM} (90% CI)	1.9 (1.37 to 2.43)	1.2 (0.48 to 1.92)	−0.34 (−0.64 to −0.04)
Smallest worthwhile change, SWC (%)	3.7	4.8	0.9
Minimal Detectable change, MDC (%)	20.2	19.5	6.0

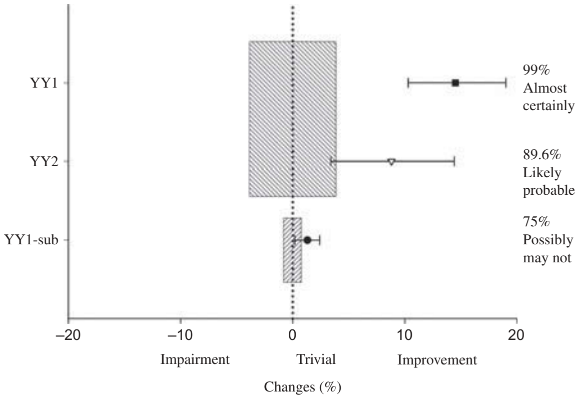


Figure 2. Percentage changes and 90% CIs in YY1, YY2 and YY1-sub and probability of the substantiality for the changes between the tests. The trivial area (grey area) was different for maximal tests (YY1 and YY2) and submaximal test (YY1-sub) and calculated from the smallest worthwhile change.

internal responsiveness in YY1-sub were assessed between T2 and T5. Two players were excluded for analysis due to loss of the data caused by a malfunction of the HR system. The ES and ES_{TEM} and 90% CIs for internal responsiveness are presented in Table II. The YY1 showed higher internal responsiveness compared to YY2 and YY1-sub.

Interchangeability

The difference in peak HR between YY1 and YY2 was 4 beats · min^{-1} (90% CI, 2, 5) in T1, 6 beats · min^{-1} (90% CI, 4, 9) in T2, 6 beats · min^{-1} (90% CI, 4, 8) in T3, 4 beats · min^{-1} (90% CI, 2, 5) in T4 and 2 beats · min^{-1} (90% CI, 1, 3) in T5, with the YY1 inducing greater values compared to YY2.

The correlation between change scores in YY1 vs. YY2 (Figure 3) and change scores in YY1 vs. respective HR6th values (i.e. YY1-sub) (Figure 4) were both small (r = 0.29 and 0.21, respectively). In addition, the correlation between change score in YY2 and HR6th was r = 0.19 (90% CI, −0.19, 0.53).

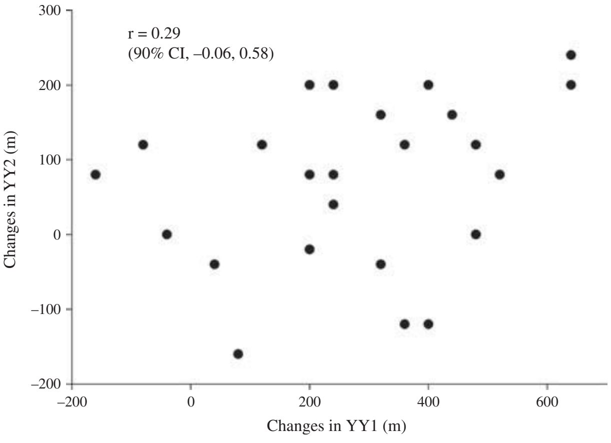


Figure 3. Interchangeability (correlation between change scores in the tests) between YY1 and YY2.

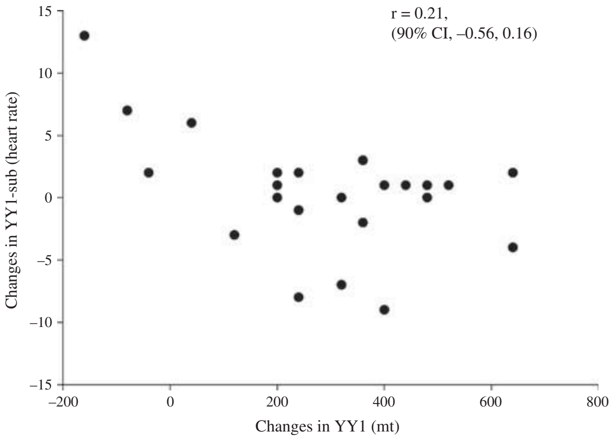


Figure 4. Interchangeability (correlation between change scores in the tests) of the YY1 and YY1-sub.

Correlations between distances reached in YY1 and YY2 were large to very large, for T1 r = 0.76 (90% CI, 0.57, 0.88), for T2 r = 0.84 (90% CI, 0.69, 0.92), for T3 r = 0.56 (90% CI, 0.27, 0.76), for T4 r = 0.66 (90% CI, 0.40, 0.82) and for T5 r = 0.69 (90% CI, 0.46, 0.84). Correlations between HR6th and the distance covered in the YY1 were moderate to very large, for T1 r = −0.40 (90% CI, −0.66, −0.86), for T2 r = −0.81 (90% CI, −0.91, −0.64), for T3 r = −0.36 (90% CI, −0.66, 0.03), for T4 r = −0.75 (90% CI, −0.87, −0.53) and for T5 r = −0.63 (90% CI, −0.80, −0.35).

Discussion

This study showed that the YYIRTs had similar reproducibility; however, the YY1 was more responsive compared to YY2. Both versions do not appear to be able to detect small but important differences at an individual level and were not interchangeable. The YY1-sub provided the higher value for absolute reliability (lower TEM) compared to maximal versions but worse ES_{TEM}, suggesting its lack in sensitivity to the changes. Furthermore, the HR6th (YY1-sub) did not reflect the changes in YY1 and YY2.

Absolute and relative reliability

The absolute reliability of the tests was similar to the values reported previously. For example, Krustrup et al. (2003, 2006) reported a CV of 4.9% and 9.6% in YY1 and YY2, respectively. Similarly, Bangsbo et al. (2008) and Thomas, Dawson, and Goodman (2006) found a CV of 8.1% and 8.7% for the YY1 and 10.4% and 12.7% for the YY2, respectively. Finally, Mohr and Krustrup (2014) recently reported a TEM expressed as CV of 12.7% to 8.6% and 1.9% to 2.3% in different periods of the season for YY2 and YY1-sub, respectively. With the exception of the studies of Thomas et al. (2006) and Mohr and Krustrup (2014), the absolute reliability has been assessed with the CV, which is considered less appropriate (Atkinson & Nevill, 1998; Hopkins, 2000) compared to the TEM as was used in the current study. However, the TEMs reported by Mohr and Krustrup (2014) could have been influenced by the specific training intervention, as suggested by significant differences found in intratrials (i.e. pre- to start-season and start- to mid-season). Therefore, a better control of training between trials (i.e. avoiding changes) is suggested for reliability studies.

The ICC values were higher for the YY2, suggesting it has more ability to discriminate among players compared with YY1 and YY1-sub. Whilst the relative reliability of the YYIRTs has been assessed in previous studies (Castagna, Manzi, Impellizzeri, Weston, & Barbero Alvarez, 2010; Ingebrigtsen et al., 2014; Mohr & Krustrup, 2014; Thomas et al., 2006), only one has examined the ICCs between different versions of YYIRTs (Thomas et al., 2006). Thomas et al. (2006) reported higher ICC for YY1 compared to YY2 (0.95 and 0.86, respectively). Unfortunately, for a better comparison of reliability between tests, the use of the same sample of participants is recommended in reliability studies (Ary, Jacobs, Razavieh, & Sorensen, 2006; Hopkins, 2000). Previous studies have reported an ICC of 0.92 for the YY1 in young players (Castagna et al., 2010) and an ICC of 0.88 to 0.82 and 0.76 to 0.84 in different periods of the season for YY2 and YY1-sub, respectively (Mohr & Krustrup, 2014) and 0.90 for the YY1-sub (Ingebrigtsen et al., 2014). Overall, our results support previous studies (Bangsbo et al., 2008), which suggested that the maximal YYIRTs have a good and similar level of reliability.

SWC and MDC

The SWCs in the present study (3.7, 4.8 and 0.9% for YY1, YY2 and YY1-sub, respectively) were similar to those calculated using data from previous studies (Bangsbo et al., 2008) where the SWC ranged between 1.2% and 8.5% for YY1, 1.5% and 7.5% for YY2 and 1.1% and 1.3% for YY1-sub (Bangsbo et al., 2008; Mohr & Krustrup, 2014). The assessment of the SWC is useful from a practical point of view, as it may be used to set minimum performance targets (±SWC) for subsequent tests. However, as the TEM of all YYIRTs was higher than the SWC, it is unlikely that YYIRTs can be used to detect smaller but important differences or individual changes. In addition, the MDC in YY1 (20.2%), YY2 (19.5%) and YY1-sub (6%) suggests a poor ability to detect substantial changes at an individual level (Terwee et al., 2007).

Changes after training and internal responsiveness

Sensitivity of a test to detect changes over a training period is an important but frequently overlooked characteristic (Impellizzeri & Marcora, 2009). As expected, both the YYIRTs changed after training and the chances of improvement (i.e. larger than the SWC) were almost certainly (99%) and likely probable (90%) in the YY1 and YY2, respectively (Figure 2). Improvements in YYIRTs (12–54%) have been reported in many studies (for review, see Bangsbo et al., 2008). Our results show a different improvement compared with Krustrup et al. (2003, 2006) both in YY1 (14.5% vs. 25% after 11 and 6 weeks, respectively) and in YY2 (8.8% vs. 27% and 42% in YY2 after 11, 4 and 8 weeks, respectively). The difference between the performance changes observed in the present study compared to previous investigations for both the YY1 (Krustrup et al., 2003) and YY2 (Krustrup et al., 2006) could be explained by several factors, including the playing level of the participants (i.e. fourth vs. first division) or differences in the training methodologies between these groups. The YY1-sub showed "possibly may not" (75%) chances of improvement and smaller changes compared to those reported by Krustrup et al. (2003) (−1.3% vs. −5%, respectively).

In the present study, the ESs of the YY1, YY2 and YY1-sub were similar to previous values reported elsewhere. Indeed, Buchheit and Rabbani (2014) and Castagna, Impellizzeri, Chaouachi, and Manzi (2013) found an ES of 1.2 and 2.1, respectively, for the YY1. Similarly, Mohr and Krustrup (2014) reported an ES of 0.5 to 1.2 and 0.4 to 1.4 calculated in different periods of the season in YY2 and YY1-sub, respectively. In order to obtain a better comparison between the tests (Husted et al., 2000), we have reported different statistics (i.e. ES and $\mathrm{ES_{TEM}}$) that have been recommended for calculating internal responsiveness. The YY1 showed higher ESs compared with YY2 and YY1-sub. The changes induced by the same training were higher for YY1 (14.5%) compared to YY2 (8.8%) and YY1-sub (−1.3%). Therefore, when reliability (noise) is compared to the changes (signal), the $\mathrm{ES_{TEM}}$ for YY1 (1.9) was superior to YY2 (1.2) and YY1-sub (−0.3). According to these results, the YY1 appears to be more useful compared to YY2 and YY1-sub for quantifying changes induced by training interventions. The changes found in our study may be limited by the specificity of the type of training imposed to the players. However, the training approach adopted with this group of players (e.g. interval training, sprint training, small-sided games, technical and tactical exercises) is typical of many soccer teams.

Interchangeability

Despite the YY1 and YY2 both being maximal tests (Bangsbo et al., 2008), our findings showed that the YY1 elicits a higher peak HR (difference from 2% to 4% between T1 and T5) compared to YY2. This finding should be considered when selecting a test to determine peak HR for training purposes. However, in contrast to the present results, several other studies have reported similar peak HR responses to both tests (Ingebrigtsen et al., 2012; Karakoç et al., 2012; Rampinini et al., 2010), suggesting that the differences between these studies could be due to the sample characteristics (age and level).

The present observation of large (r = 0.56) to very large (r = 0.84) correlations between performance in YYIRTs is in accordance with several previous studies (Ingebrigtsen et al., 2012, 2014; Mohr & Krustrup, 2014; Rampinini et al., 2010; Thomas et al., 2006). For example, both Ingebrigtsen et al. (2012, 2014) and Mohr and Krustrup (2014) found very large correlations between YY1 and YY2 distances (r = 0.74, r = 0.76, r = 0.75 and r = 0.77, respectively), whilst Rampinini et al. (2010) and Thomas et al. (2006) reported large correlations between the two tests (r = 0.70 and r = 0.50 to 0.63, respectively). However, in contrast to these results, Karakoç et al. (2012) revealed a nonsignificant correlation (r = 0.52) between tests. This difference may have been due to the lower sample size (n = 12) used in the previous study.

To the best of our knowledge, this is the first study in which the correlation between change scores in the YYIRTs has been assessed. If the two tests are interchangeable, the change scores (distances) in YY1 should reflect the change scores in the YY2 or in YY1-sub (HR). The results of the present study showed that only 8% of variance was explained by changes between the two levels of the YYIRT, highlighting that the two tests are not interchangeable. Furthermore, we found that even the YY1-sub was not interchangeable with YY1 and YY2 (4% and 3% of the variance explained, respectively). Therefore, the YYIRTs should be considered independently and we recommend that they be selected according to the aim of the assessment and the different physiological responses targeted (Rampinini et al., 2010).

Conclusions

The present study showed that the reliability of both maximal YYIRTs were similar and lower than the training-induced changes for each test, demonstrating adequate responsiveness at a group level. However, the acceptability of the TEMs (noise) depends on the magnitude of changes (signal), and for this reason, the reliability of YY1 was better than that of YY2 for detecting training-induced changes. However, despite the YY1 being more responsive to training, the YY2 is a shorter test and therefore may be preferred by players over the YY1. This study also showed that although the two maximal YYIRTs are correlated, they measure different physical characteristics (low convergent validity), and therefore, they are not interchangeable. Whilst both tests could be used, the present results show that the YY1 provides more useful information. Due to its submaximal intensity and short duration, the YY1-sub could be useful for the physical assessment during rehabilitation or regular assessment of a player's fitness during the competition season. However, this test appears to have poorer sensitivity for detecting the training-induced effects compared to YY1. In conclusion, the YY1 demonstrated important test characteristics such as construct validity, reliability and responsiveness. However, a definitive confirmation of the validity of this test would require an examination of the correlation between change scores in the test and changes in high-intensity activity (external responsiveness) performed during a match or using controlled match simulations. Unfortunately, to our knowledge, this kind of validation has not yet been verified. Therefore, further studies in this direction are warranted.

References

Aaronson, N., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D. L., Perrin, E., & Stein, R. E. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research*, *11*(3), 193–205.

Amann, M., Hopkins, W. G., & Marcora, S. M. (2008). Similar sensitivity of time to exhaustion and time-trial time to changes in endurance. *Medicine & Science in Sports & Exercise*, *40*(3), 574–578.

Ary, D., Jacobs, L. C., Razavieh, A., & Sorensen, C. (2006). *Introduction to Research in Education*. Belmont, CA: Wadsworth.

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, *26*(4), 217–238.

Bangsbo, J., Iaia, F. M., & Krustrup, P. (2008). The Yo-Yo intermittent recovery test: A useful tool for evaluation of physical performance in intermittent sports. *Sports Medicine*, *38*(1), 37–51.

Batterham, A. M., & Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *International Journal of Sports Physiology and Performance*, *1*(1), 50–57.

Beckerman, H., Roebroeck, M. E., Lankhorst, G. J., Becher, J. G., Bezemer, P. D., & Verbeek, A. L. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research*, *10*(7), 571–578.

Buchheit, M., & Rabbani, A. (2014). 30–15 intermittent fitness test vs. Yo-Yo intermittent recovery test level 1: Relationship and sensitivity to training. *International Journal of Sports Physiology and Performance¸ 9*(3), 522–524.

Castagna, C., Impellizzeri, F. M., Chaouachi, A., & Manzi, V. (2013). Preseason variations in aerobic fitness and performance in elite-standard soccer players: A team study. *Journal of Strength and Conditioning Research*, *27*(11), 2959–2965.

Castagna, C., Manzi, V., Impellizzeri, F., Weston, M., & Barbero Alvarez, J. C. (2010). Relationship between endurance field tests and match performance in young soccer players. *Journal of Strength and Conditioning Research*, *24*(12), 3227–3233.

de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health Quality of Life Outcomes*, *4*, 54.

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–1039.

Ferrari Bravo, D., Impellizzeri, F. M., Rampinini, E., Castagna, C., Bishop, D., & Wisloff, U. (2008). Sprint vs. interval training in football. *International Journal of Sports Medicine*, *29*(8), 668–674.

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

Hopkins, W. G. (2007). A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a P value. *Sportscience*, *11*, 16–20.

Husted, J. A., Cook, R. J., Farewell, V. T., & Gladman, D. D. (2000). Methods for assessing responsiveness: A critical review and recommendations. *Journal of Clinical Epidemiology*, *53*(5), 459–468.

Impellizzeri, F. M., & Marcora, S. M. (2009). Test validation in sport physiology: Lessons learned from clinimetrics. *International Journal of Sports Physiology and Performance*, *4*(2), 269–277.

Impellizzeri, F. M., Marcora, S. M., Castagna, C., Reilly, T., Sassi, A., Iaia, F. M., & Rampinini, E. (2006). Physiological and performance effects of generic versus specific aerobic training in soccer players. *International Journal of Sports Medicine*, *27* (6), 483–492.

Impellizzeri, F. M., Rampinini, E., Castagna, C., Bishop, D., Ferrari Bravo, D., Tibaudi, A., & Wisloff, U. (2008). Validity of a repeated-sprint test for football. *International Journal of Sports Medicine*, *29*(11), 899–905.

Ingebrigtsen, J., Bendiksen, M., Randers, M. B., Castagna, C., Krustrup, P., & Holtermann, A. (2012). Yo-Yo IR2 testing of elite and sub-elite soccer players: Performance, heart rate response and correlations to other interval tests. *Journal of Sports Sciences*, *30*(13), 1337–1345.

Ingebrigtsen, J., Brochmann, M., Castagna, C., Bradley, P. S., Ade, J., Krustrup, P., & Holtermann, A. (2014). Relationships between field performance tests in high-level soccer players. *Journal of Strength and Conditioning Research*, *28*(4), 942–949.

Karakoç, B., Akalan, C., Alemdaroğlu, U., & Arslan, E. (2012). The relationship between the Yo-Yo tests, anaerobic performance and aerobic performance in young soccer players. *Journal of Human Kinetics*, *35*, 81–88.

Krustrup, P., Mohr, M., Amstrup, T., Rysgaard, T., Johansen, J., Steensberg, A., … Bangsbo, J. (2003). The Yo-Yo intermittent recovery test: Physiological response, reliability, and validity. *Medicine & Science in Sports & Exercise*, *35*(4), 697–705.

Krustrup, P., Mohr, M., Nybo, L., Jensen, J. M., Nielsen, J. J., & Bangsbo, J. (2006). The Yo-Yo IR2 test: Physiological response, reliability, and application to elite soccer. *Medicine & Science in Sports & Exercise*, *38*(9), 1666–1673.

Liow, D. K., & Hopkins, W. G. (2003). Velocity specificity of weight training for kayak sprint performance. *Medicine & Science in Sports & Exercise*, *35*(7), 1232–1237.

Mohr, M., & Krustrup, P. (2014). Yo-Yo intermittent recovery test performances within an entire football league during a full season. *Journal of Sports Sciences*, *32*(4), 315–327.

Mujika, I., Santisteban, J., Impellizzeri, F. M., & Castagna, C. (2009). Fitness determinants of success in men's and women's football. *Journal of Sports Sciences*, *27*(2), 107–114.

Norman, G. R., Wyrwich, K. W., & Patrick, D. L. (2007). The mathematical relationship among different forms of responsiveness coefficients. *Quality of Life Research*, *16*(5), 815–822.

Rampinini, E., Impellizzeri, F. M., Castagna, C., Abt, G., Chamari, K., Sassi, A., & Marcora, S. M. (2007). Factors influencing physiological responses to small-sided soccer games. *Journal of Sports Sciences*, *25*(6), 659–666.

Rampinini, E., Sassi, A., Azzalin, A., Castagna, C., Menaspà, P., Carlomagno, D., & Impellizzeri, F. M. (2010). Physiological determinants of Yo-Yo intermittent recovery tests in male soccer players. *European Journal of Applied Physiology*, *108*(2), 401–409.

Terwee, C. B., Bot, S. D., De Boer, M. R., Van Der Windt, D. A., Knol, D. L., Dekker, J., … de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42.

Thomas, A., Dawson, B., & Goodman, C. (2006). The Yo-Yo test: Reliability and association with a 20-m shuttle run and VO (2max). *International Journal of Sports Physiology and Performance*, *1*(2), 137–149.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231–240.