**Criterion-Related Validity of the 20-M Shuttle Run Test for Estimating Cardiorespiratory Fitness: A Meta-Analysis**

Daniel Mayorga-Vega, Pablo Aguilar-Soto and Jesús Viciana
Department of Physical Education and Sport, University of Granada, Spain

**Abstract**

The main purpose of the present meta-analysis was to examine the criterion-related validity of the 20-m shuttle run test for estimating cardiorespiratory fitness. Relevant studies were searched from twelve electronic databases up to December 2014, as well as from several alternative modes of searching. The Hunter-Schmidt’s psychometric meta-analysis approach was conducted to estimate the population criterion-related validity of the 20-m shuttle run test. From 57 studies that were included in the present meta-analysis, a total of 78 correlation values were analyzed. The overall results showed that the performance score of the 20-m shuttle run test had a moderate-to-high criterion-related validity for estimating maximum oxygen uptake ($r_p = 0.66-0.84$), being higher when other variables (e.g. sex, age or body mass) were used ($r_p = 0.78-0.95$). The present meta-analysis also showed that the criterion-related validity of Léger’s protocol was statistically higher for adults ($r_p = 0.94, 0.87-1.00$) than for children ($r_p = 0.78, 0.72-0.85$). However, sex and maximum oxygen uptake level do not seem to affect the criterion-related validity values. When an individual’s maximum oxygen uptake attained during a laboratory-based test is not feasible, the 20-m shuttle run test seems to be a useful alternative for estimating cardiorespiratory fitness. In adults the performance score only seems to be a strong estimator of cardiorespiratory fitness, in contrast among children the performance score should be combined with other variables. Nevertheless, as in the application of any physical fitness field test, evaluators must be aware that the performance score of the 20-m shuttle run test is simply an estimation and not a direct measure of cardiorespiratory fitness.

**Key words:** Maximum oxygen uptake, peak oxygen uptake, PACER, Multistage fitness test, Léger test.

**Introduction**

Nowadays, cardiorespiratory fitness is considered one of the most powerful markers of health, even above other traditional markers such as blood pressure or cholesterol level (Blair, 2009). Current evidence has shown how cardiorespiratory fitness status is an important quantitative predictor of cardiovascular events and all-cause mortality in healthy adults (Kodama et al., 2009). Additionally, during childhood higher cardiorespiratory fitness levels have been associated with a healthier cardiovascular profile in adulthood (Ruiz et al., 2009). Therefore, cardiorespiratory fitness testing may help to identify a target population for primary prevention both in children and adults, as well as for health promotion policies (Ruiz et al., 2009).

Different kinds of tests are commonly used to assess cardiorespiratory fitness. Cardiorespiratory fitness is typically identified as the maximal oxygen uptake ($VO_2_{max}$) reached by an individual (Pescatello et al., 2014). Specifically, the $VO_2_{max}$ attained during a laboratory-based and graded maximal exercise test is widely considered the criterion measure (also called “gold standard”) of cardiorespiratory fitness (Pescatello et al., 2014). Alternatively, due to advances in technology, today a portable gas analyzer can also be worn during a field-based graded maximal exercise test (Castagna et al., 2010; Silva et al., 2012). Due to the necessity of sophisticated and costly instrumentation, qualified technicians, and time constraints, the use of the directly measured $VO_2_{max}$ is limited in several settings such as in sports clubs, schools, or in large scale research studies (Pescatello et al., 2014).

Unlike the direct methods to determine $VO_2_{max}$, in the above mentioned settings the performance score attained during cardiorespiratory fitness field tests could be a useful alternative. The 20-m shuttle run test (20MSR) test, also called the “Course Navette”, “PACER”, or “Multistage fitness test”, is probably the most widely used field test for estimating cardiorespiratory fitness (Castro-Piñero et al., 2010). The 20MSR test is simple, easy to administer and not too time-consuming, it requires minimal equipment, and a large number of individuals can be tested simultaneously. The 20MSR test consists of one-minute stages of continuous, incremental speed running. The initial speed is 8.5 km/h, and increases by 0.5 km/h per minute (Léger et al., 1984). The individual is required to run between two lines 20-m apart, while keeping pace with audio signals emitted from a pre-recorded cassette or compact disk. The test ends when the individual fails to reach the end lines concurrent with the audio signals on two consecutive occasions. Although in the original protocol the stages lasted two minutes (Léger and Lambert, 1982), later it was modified to one minute stages which were considered more motivating (Léger et al., 1984). Additionally, later different combinations of starting speed and speed increase have been proposed (e.g. Cadenas-Sánchez et al., 2014; Dong-Ho et al., 2014).

Each primary study that is published about criterion-related validity of the 20MSR test only constitutes a single piece of a constantly growing body of evidence (Cooper et al., 2009). For instance, in some studies the correlation coefficient is high (Chatterjee et al., 2006c), while in others the association is moderate or even low (Von Haaren et al., 2011). To make sense of the often conflicting results found in the scientific literature, researchers have to conduct meta-analyses (Cooper et al., 2009; Hunter and Schmidt, 2004; Lipsey and Wilson,
2001). Thus meta-analyses remain a useful tool for the evaluation of evidence (Cooper et al., 2009), forming a critical process for the development of theory in science (Hunter and Schmidt, 2004).

Previous studies have carried out meta-analyses on the validity of different field-based tests widely used in sports sciences such as the Borg’s perceived exertion scale (Chen et al., 2002), the International Physical Activity Questionnaire (Kim et al., 2012), or the flexibility tests sit-and-reach (Mayorga-Vega et al., 2014a) and toe-touch (Mayorga-Vega et al., 2014b). To our knowledge there are not any meta-analyses addressing the criterion-related validity of the 20MSR test. Therefore, the purposes of the present meta-analysis were: (a) to estimate and compare the overall population mean of the criterion-related validity coefficients of the 20MSR test for estimating cardiorespiratory fitness; (b) to examine the influence of some study features (sex, age, and level of VO2max of the participants) in criterion-related validity coefficients of the 20MSR test (between-study analyses); and (c) to compare the values of the criterion-related validity coefficients between the performance score only and the performance score combined with other variables (within-study analyses).

Methods

Search strategy

The following twelve electronic bibliographic databases were searched through December 2014: Web of Science (all databases), Scopus, SportDiscus, CINAHL, Cochrane Library Plus, ERIC, ProQuest Education Journals, Applied Social Sciences Index and Abstracts, ProQuest Social Science Journals, International Bibliography of the Social Sciences, Proquest Dissertations and Theses, and WorldCat. The searches were carried out in the search field type “Title, abstract, and keywords” or equivalent (e.g. “Topic” for the Web of Science database). Any publication format including journal papers and grey literature (i.e. master/doctoral dissertations and conference proceedings) was examined. Additionally, no language or publication date restrictions were imposed.

The search terms used were based on two concepts. Concept one included terms for the 20MSR test (navette, Léger, shuttle run, shuttle-run, shuttle test*, shuttle endurance run, multistage fitness, multi-stage fitness, beep, beep, progressive aerobic cardiovascular endurance, PACER, 20 m test*, 20-m test*, 20 m run, 20-m run, bip test*) and concept two included terms related to validity (valid*, related, relationship, correlation, regression, comparison, association, estimat*, determinat*, predict*, equation*, VO2*, oxygen uptake, oxygen intake, consumption of oxygen, oxygen consumption, aerobic, cardiovascular, cardiorespiratory, fitness, gold standard, criterion measure*). The truncated root of certain terms was followed by an asterisk to include multiple variants. Additionally, the keywords that consisted of more than one word were enclosed in quotes. Finally, the terms of the same concept were combined together with the Boolean operator “OR” and then the two concepts were combined using the Boolean operator “AND” (Cooper et al., 2009).

Based on the results of the Boolean-based search (as well as all the related studies by Léger), other modes of searching were carried out. The reference lists of all studies (as well as some related studies reviews) were manually searched (also called “snowballing”). Additionally, the reference citations (in the Web of Science and Scopus databases) and the researcher publications of the first authors (in the Web of Science, Scopus and SportDiscus databases) were also examined. Subsequently, the authors for correspondence (if they were not defined, the first author was used) were contacted by email. Finally, the researcher’s personal lists (in ResearchGate, Google Scholar, and personal websites) of the first authors were screened. Any time a new record was found, all of these modes of searching were repeated until any new study appeared.

Selection criteria

The selection criteria to identify studies that examined the criterion-related validity of the 20MSR test were: (1) studies with apparently healthy participants who did not present any injury, physical and/or mental disabilities; (2) studies with the original protocols of the 20MSR test (Léger and Lambert, 1982; Léger et al., 1984) or some modifications of them in starting speed, speed increase and/or duration of stages (also called “levels” or “paliers”); (3) studies in which for the criterion measure the VO2max was measured in a standardized and laboratory-based incremental test to exhaustion; and (4) studies which associated the performance scores of the field test (or the performance score with other variables) with the measured VO2max results using a Pearson’s r zero-order correlation coefficient or simple linear regression ($R^2$) (or a multiple linear regression in case of multiple predictors).

Coding studies

For the present meta-analysis, from each selected study the following data were coded: Identification number, type of publication (1 = journal paper, 2 = grey literature –master dissertation, doctoral dissertation, or conference proceeding-), sample size ($n$), sex of participants (1 = men, 2 = women, 3 = men and women, 4 = no information), age of participants (1 = children, < 18 years; 2 = adults, ≥ 18 years; 3 = children and adults; 4 = no information), 20MSR test protocol (1 = Léger’s protocol, 2 = Eurofit protocol, 3 = QUB’s protocol; 4 = others), criterion measure protocol (1 = treadmill run test; 2 = cycle ergometer test; 3 = others), measurement unit of the 20MSR test (1 = completed stage, accuracy of one; 2 = completed stage, accuracy of half; 3 = total laps; 4 = speed expressed in km/h; 5 = time expressed in seconds; 6 = distance expressed in metres; 7 = others), measurement unit of the criterion test (1 = VO2max expressed in ml/kg/min; 2 = VO2max expressed in l/min; 3 = maximal aerobic speed expressed in km/h; 4 = other), mean value of the measurement criterion, reliability of the 20MSR test (intraclass correlation coefficient), reliability of the measurement criterion (intraclass correlation coefficient), statistical test used for the criterion-related validity (1 =
Pearson’s $r$ correlation coefficient, $2 = R^2$ simple or multiple linear regression), and criterion-related validity value (separately for performance score only and multiple predictors). In addition, any observations were also registered when some special question was found.

Although various protocols for evaluating study quality have been described, there is no widespread agreement on the validity of this kind of evaluation approach (e.g. see Cooper et al., 2009). Thus, rejecting certain studies and accepting others for inclusion in a meta-analysis on the basis of a quality score remains a controversial procedure (Cooper et al., 2009; Flather et al., 1997). Therefore, according to Flather et al. (1997), in the present meta-analysis the approach followed has been to ensure that the design has not been flawed (e.g. the $VO_2$ max was measured in a standardized and laboratory-based incremental test to exhaustion), and that there has been a complete reporting of relevant outcomes. For a study to be included in this meta-analysis, sample size, protocol of the 20MSR test, unit and protocol of the criterion measure test, statistical test, and value of the criterion-related validity of the 20MSR test (Hunter and Schmidt, 2004) were considered the unit of measurement as an indication of the criterion-related validity of the 20MSR test. When the validity values were reported as $R^2$, therefore, it was transformed. Additionally, to avoid dependency issues in the meta-analysis, an exhaustive examination of the selected studies was carried out. All the examined studies used the relative $VO_2$ max (i.e. expressed in ml/kg/min) as the measurement criterion. Although some studies also reported criterion-related validity results using additional markers such as the absolute $VO_2$ max expressed in l/min (Aziz et al., 2007; McIver et al., 2004), relative $VO_2$ max using lean body mass (Varness et al., 2009), or the maximal aerobic speed (Kuisis, 2007), these validity coefficients were not selected. Since some studies used multiple performance scores of the 20MSR test for examining the criterion-related validity (LaMontagna, 1991; Matsuzaka et al., 2004; Metsios et al., 2006; Ramsbottom et al., 1988; Suminski et al., 2004; Varness et al., 2009), the average value was used. Nevertheless, when authors reported the results of criterion-related validity from the combination of different multiple predictors (Barnett et al., 1993; Mahar et al., 2006; 2011; Hamlin et al., 2014), only the best model (i.e. higher coefficient value) was used in the present meta-analysis.

If a single study reported more than one $r$ value within the same 20MSR test protocol, but from different subsamples, we assumed each $r$ value from different subsamples to be independent and included them in a single meta-analysis (Lipsy and Wilson, 2001). When, in the same study, data for men and women were expressed both separately and together, only the separate data were selected (e.g. Hamlin et al., 2014; Silva et al., 2012; Von Haaren et al., 2011). However, when in the same study, data for the whole and subsamples with respect to sex and age categories were expressed, only the whole sample was coded (e.g. Mahar et al., 2006). Similarly, when in the same study, data were expressed for different days from the same sample (i.e. LaMontana, 1991; McIver et al., 2004), the average value of the coefficients was coded.

**Publication bias:** In addition to the search strategy followed and selection criteria to avoid availability bias, another examination of the selected studies was carried out to avoid a potential duplication of information retrieved. Similarities between studies of the same authors, with the same correlation coefficients and/or the same sample size were examined. If some selected studies had full or partial duplicated information, these particular correlation values were not analyzed in the meta-analyses. Furthermore, before computing correlations, several exploratory analyses were also conducted for identifying and assessing the impact of any potential publication bias. Firstly, according to Light and Pillemer (1984), the scatter plots of correlation coefficients against sample size for each 20MSR test protocol were analyzed. Secondly, with the objective of quantifying the outcomes of the scatter plots, based on Begg and Mazumdar (1994), a Spearman’s rank order correlation between $r$ values and sample size was calculated. Finally, for assessing the impact of any potential publication bias, a file drawer analysis based on effect size was performed to estimate the number of unlocated studies averaging null results ($r = 0$) that would have to exist to bring the mean effect size ($r_p$) down to the small mean $r$ value (Orwin, 1983). According to Cohen’s (1992) guidelines, the correlation coefficient was interpreted as small when $r < 0.30$.

**Computation of correlations:** The Hunter-Schmidt’s psychometric meta-analysis approach was conducted to obtain the population estimates of the criterion-related validity of the 20MSR test (Hunter and Schmidt, 2004). This approach estimates the population correlation correcting the observed correlations due to various artefacts such as sampling error and measurement error. The “bare-bone” mean $r$ ($r_c$), corrected for only sampling error was first calculated by weighting each $r$ with the respective sample size when aggregating them into $r_c$. Then, we calculated the corrected mean $r$ at the population level ($r_p$) that was unaffected by both sampling error and measurement error. Since the reliability coefficients (intraclass correlation coefficients) of the 20MSR were unavailable in most of the included primary studies, the measurement error was corrected using artifact distributions instead of individually. On the other hand, the
measurement error of the criterion test could not be corrected because the reliability was not available. Finally, the 95% confidence intervals of $r_p$ (95% CI) were calculated.

**Moderator analysis:** In the present meta-analysis, due to the low number of $r$ values found, partial hierarchical analyses of moderator variables were carried out. According to Hunter and Schmidt (2004), to determine the presence of moderator effects which may affect overall criterion-related validity of the 20MSR test ($r_p$), three different criteria were simultaneously examined: (a) the 95% credibility interval (95% CV) is relatively large or includes the value zero; (b) the percentage of variance accounted for by statistical artefacts is less than 75% of the observed variance in $r_p$; and (c) the Q homogeneity statistic is statistically significant ($p < 0.05$). If at least one of the three criteria were met, we concluded that the results could be affected by moderator effects. In the presence of moderator effects, criterion-related validity values of the 20MSR test were analyzed separately by: (a) sex of participants (i.e. men and women); (b) age of participants (i.e. children and adults); and (c) level of VO$_2$max (i.e. low average level, $< P_{50}$, and high average level, $\geq P_{50}$) (between-study analyses). Additionally, the criterion-related validity values of the 20MSR test for the performance score only were compared with the criterion-related validity with multiple predictors (within-study analysis).

The meta-analyses were performed using the software Hunter and Schmidt Meta-Analysis Programs version 1.1 for Windows (Iowa, 2005). All the other statistical analyses were performed using the SPSS version 20.0 for Windows (IBM® SPSS® Statistics 20).

**Results**

**Study description**

Figure 1 shows a flowchart of the study selection process. Of the 7,777 bibliographic databases search results, 238 potentially relevant publications were identified and retrieved for a more detailed evaluation. Afterward, based on the 52 studies of the Boolean-based search that met the selection criteria (plus 16 studies reviews that were also used for the reference lists mode), other modes of searching were carried out. Through the other modes of searching eight additional studies met the selection criteria. However, due to duplication, of the overall 60 studies that met the inclusion criteria, 57 studies were included in the present meta-analysis. Finally, from the 57 studies that were included in the present meta-analysis, a total of 78 $r$ values across three 20MSR test protocols were retrieved, being 65 $r$ correlation coefficients for the criterion-related validity using the performance score only and 13 $r$ correlation coefficients for multiple predictors (i.e. the performance score and other variables: age, sex, biological maturation, body mass, body mass index, body fat and/or skinfolds).

In the present meta-analysis 54 studies with performance score only (Aandstad et al., 2011; Armstrong et al., 1988; Aslan et al., 2012; Aziz et al., 2005b; 2007; Bandyopadhyay, 2011; 2013; Barnett et al., 1993; Chatterjee et al., 2005; 2006a; 2006c; 2008a; 2008b; 2008c; 2009; 2010a; 2010b; 2010c; 2011; 2013; De Souza et al., 2010; Dickau, 2011; Dong-Ho et al., 2014; Flouris et al., 2004; 2006; Gadoury and Léger, 1986; Green et al., 2013; Hamlin et al., 2014; Kuisis, 2007; LaMontagna, 1991; Liu et al., 1992; Mahar et al., 2002; 2006; 2011; 2013; Mahoney, 1992; Matsuzaka et al., 2004; McIver et al., 2004; McVeigh et al., 1995; Metsios et al., 2006; Mombiedro et al., 1992; O’Gorman et al., 2000; Palizcka et al., 1987; Paradisis et al., 2014; Pitetti et al., 2002; Poortmans et al., 1986; Ramsbottom et al., 1988; Stickland et al., 2003; Suminski et al., 2004; Thomas et al., 2006; Van Mechelen et al., 1986; Van Praagh et al., 1988; Varness et al., 2009; Von Haaren et al., 2011) and 11 studies with multiple predictors were included (Barnett et al., 1993; Chia et al., 2005; Dong-Ho et al., 2014; Hamlin et al., 2014; Mahar et al., 2002; 2006; 2010; 2011; McVeigh et al., 1995; Matsuzaka et al., 2004; Tsiaras et al., 2010).

Some studies retrieved for a more detailed evaluation were not included because they were carried out with non-healthy participants (e.g. individuals with Down’s syndrome, cerebral palsy or in wheelchairs) (e.g. Agiovlasitis et al., 2011; Goosey-Tolfrey and Tolfrey, 2008; Kloyiam et al., 2011), used major modifications of the 20MSR test (e.g. the Square shuttle run test or Yo-Yo intermittent recovery test) (e.g. Castagna et al., 2008). Other studies that were retrieved for a more detailed evaluation were not selected because only cross-validity was examined (e.g. Batista et al., 2013). Not one of three potential studies were selected because they did not define the protocol used (i.e. lacked critical information) and the authors did not reply when asked for it (Cunningham et al., 1994; Hemmings et al., 2003; Lightburne, 2008). Then, the full-text of some potential studies was not available (Barnejee et al., 2005; Chartterjee et al., 2007).

Finally, some potential studies were not selected because they did not use the measured VO$_2$max during a standardized and laboratory-based incremental test to exhaustion as a criterion measure. For instance, some research studies assessed the VO$_2$max during the field test (e.g. Castagna et al., 2010; Silva et al., 2012). Nevertheless, previous studies have found that the measured VO$_2$max during the 20MSR test is significantly different compared with that measured during a laboratory-based test (Aziz et al., 2005a; Flouris et al., 2010). In other potential studies (e.g. Léger and Lambert, 1982; Léger et al., 1988) the VO$_2$max was assessed by retroextrapolating the O$_2$ recovery curve at time zero of recovery. Retroextrapolation is a method to estimate VO$_2$max (i.e. indirect measure) and, therefore, it cannot be considered as a criterion measure to determine it (i.e. direct measure) such as assessing the VO$_2$max during a standardized incremental test to exhaustion (Aslan et al., 2012; Mahar et al., 2011).

**Publication bias**

Firstly, several exploratory analyses were followed to avoid full or partial duplicated information availability bias. Although three research studies met the selection criteria (Chartterjee et al., 2006b; Hamlin et al., 2013; Paradisis et al., 2013), the correlation coefficient value was not analyzed in the present meta-analyses. Paradisis
et al. (2013) and Hamlin et al. (2013) conference papers were not included because they were published later in a journal (Hamlin et al., 2014; Paradisis et al., 2014). Actually, Paradisis and colleagues (2013) in their conference paper published the results of a pilot study using a subsample. Regarding the study by Chatterjee et al. (2006b), the same study had been also published in another journal (Chatterjee et al., 2008c).

Exploratory analyses were conducted to identify the presence of publication bias. Because the sum of the $r$ values for some protocols was very small, the following analyses were calculated only for Léger’s (for both performance score only and multiple predictors) and Eurofit protocols (for performance score only). Figure 2 shows the scatter plots of sample size against criterion-related validity coefficients for estimating VO2max for Léger’s (for performance score only and multiple predictors) and Eurofit protocols (for performance score only). According to this graphical method, the figures suggest that for the performance score only there was not publication bias for both protocols. Nevertheless, we have to be aware that for Léger’s protocol seems to be a slightly major density in the right-hand corner. For Léger’s protocol with multiple predictors, however, the scatter plot suggests the presence of publication bias, because of the absence of $r$ values in the lower left hand corner.

The results of Spearman’s rank order correlation between $r$ values and sample size did not show any statistically significant correlation (Léger’s protocol, performance score only: $r = 0.08$, $p = 0.601$; multiple predictors: $r = -0.26$, $p = 0.450$; Eurofit protocol, $r = -0.06$, $p = 0.873$). Due to the small number of $r$ values found for the Léger’s protocol (with multiple predictors) and the Eurofit protocol, the results of both the scatter plot and the Spearman correlation must be interpreted with caution (Begg and Mazumdar, 1994; Cooper et al., 2009). Addi-
tionally, empirical evaluations of the funnel plots suggest that their interpretation can be limited (Terrin et al., 2005).

Figure 2. Scatter plot of sample size against criterion-related validity coefficients ($r$) for estimating maximal oxygen uptake: (a) Léger’s protocol with performance only score; (b) Léger’s protocol with multiple predictors; and (c) Eurofit protocol with performance score only. Dashed line represents mean values of validity coefficients.

Finally, file drawer analyses based on effect size were carried out for assessing the impact of any potential publication bias. The results of the file drawer analyses are based on effect size for estimating the number of unlocated studies averaging null results ($r = 0$) that would have to exist to bring the mean $r_p$ down to 0.29. These results are shown in the following lines (in parenthesis the unlocated/located percentage): for performance score only, Léger’s protocol 80 (167%), Eurofit protocol 14 (127%), QUB’s protocol 7 (140%), and Do-Hong’s protocol 2 (200%); for multiple predictors, Léger’s protocol 20 (182%), Eurofit protocol 2 (200%), and Do-Hong’s protocol 2 (200%). Although we are aware that there is not a large number of “lost” studies for some protocols, the results for the percentage of unlocated/located studies showed an unlikely number of “lost” studies (127-200%).

Criterion-related validity

Table 1 reports the number of $r$ values studied ($K$), the total sample size accumulated ($N$), the overall weighted mean of $r$ corrected for sampling error only ($r_\text{c}$), the overall weighted mean of $r$ corrected for both sampling error and measurement error ($r_\text{p}$), as well as the 95% CI for overall criterion-related validity correlation coefficients ($r_\text{p}$) for estimating VO$_2$max across each 20MSR protocol. Additionally, to detect the presence of moderator effects which may affect overall criterion-related validity of the 20MSR test, the 95% CV, the percentage of variance accounted for by statistical artefacts, and the $Q$ homogeneity statistic were calculated.

The overall results showed that the 20MSR test had a moderate-to-high mean correlation coefficient of criterion-related validity for estimating VO$_2$max in which all 95% CI did not include the value zero. The results of the present meta-analysis also showed that the criterion-related validity of Léger’s protocol was statistically higher than the QUB’s (Queen’s University Belfast) protocol. For Léger’s and Eurofit protocols the percentage of variance accounted for by statistical artefacts was less than 75%, and the 95% CV was relatively large, as well as for Léger’s protocol the $Q$ homogeneity statistic was also statistically significant ($p < 0.05$). Therefore, follow-up moderator analyses were conducted using predefined moderators as it was hypothesized in the present study. However, since none of the three criteria were met in the QUB’s protocol, moderator analyses were not conducted for that particular protocol.

Regarding the multiple predictors, the overall results showed that when the performance score of the 20MSR test was combined with other variables the mean correlation coefficients of criterion-related validity for estimating VO$_2$max were moderate-to-very-high. Additionally, when the 95% CI could be calculated (i.e. for Léger’s protocol), the value zero was not included. Although two of the three criteria were met in Léger’s protocol, due to the small $n$ for the most of its subcategories (e.g. only one correlation coefficient for adults subcategory and two for men subcategory), the between-study moderator analyses were not conducted in that case. However, due to the fact that for Léger’s protocol eight studies reported correlation coefficients of criterion-related validity for both performance score only and combined with other variables, the within-study analysis was conducted as it was hypothesized in the present study (see moderator analyses).

Moderator analyses

Table 2 shows the results of between-study moderator analyses to examine the effects of sex (i.e. men and women), the age of participants (i.e. children and adults), and
estimating VO$_2$max for both men and women in which all
correlation coefficient of criterion-related validity for
validity correlation coefficients for estimating VO$_2$max
hand, we must point out that, according to moderator
mean correlation coefficients overlapped. On the other
protocols.

dLéger’s protocol starts at 8.5 km/h and increases 0.5 km/h each minute (Léger et al., 1984, 1988); Eurofit protocol starts at 8.0 km/h and increases 0.5 km/h each minute, but the second stage increases by 1.0 km/h (Council of Europe Committee for the Development of Sport, 1988); QUB’s protocol starts at 8.0 km/h and increases 0.5 km/h each minute (Riddoch, 1990); Dong-Ho’s protocol starts at 7.5 km/h and increases 0.5 km/h each minute (Dong-Ho et al., 2014). * p < 0.05

| Table 1. Results of meta-analyses for overall criterion-related validity correlation coefficients across the 20-m shuttle run test protocols. |
| Protocols | K | N  | $r_c$ | $r_p$ | 95% CI$^a$ | 95% CV$^b$ | % variance$^c$ | Q statistic |
| Performance score only |
| Léger$^d$ | 48 | 2,222 | .77 | .84 | .80-.89 | .54-1.00 | 15.15 | 313.42* |
| Eurofit$^e$ | 11 | 278 | .66 | .73 | .60-.86 | .43-1.00 | 42.50 | 18.10 |
| QUB$^f$ | 5 | 401 | .68 | .71 | .64-.77 | .64-.77 | 79.86 | 1.34 |
| Dong-Ho$^g$ | 1 | 127 | .62 | .66 | - | - | - | - |

| Performance score with other variables |
| Léger$^h$ | 11 | 893 | .80 | .87 | .81-.94 | .67-1.00 | 14.94 | 73.04* |
| Eurofit$^i$ | 1 | 55 | .85 | .95 | - | - | - | - |
| Dong-Ho$^j$ | 1 | 127 | .73 | .78 | - | - | - | - |

Note: K, number of rs; N, total sample size; $r_c$, overall weighted mean of $r$ corrected for sampling error only; $r_p$, overall weighted mean of $r$ corrected for sampling error and measurement error of the 20-m shuttle run test; 95% CI, 95% confidence interval; 95% CV, 95% credibility interval. Percentage of variance accounted for by statistical artefacts including sampling error and measurement error of the 20-m shuttle run test; *Percentage of variance accounted for by statistical artefacts including sampling error and measurement error of the 20-m shuttle run test; †Léger’s protocol starts at 8.5 km/h and increases 0.5 km/h each minute (Léger et al., 1984, 1988); Eurofit protocol starts at 8.0 km/h and increases 0.5 km/h each minute, but the second stage increases by 1.0 km/h (Council of Europe Committee for the Development of Sport, 1988); QUB’s protocol starts at 8.0 km/h and increases 0.5 km/h each minute (Riddoch, 1990); Dong-Ho’s protocol starts at 7.5 km/h and increases 0.5 km/h each minute (Dong-Ho et al., 2014). * p < 0.05

| Table 2. Results of moderator analyses for criterion-related validity correlation coefficients across the 20-m shuttle run test protocols potentially affected by moderator effects$^k$. |
| Moderator | Effect | K | N  | $r_c$ | $r_p$ | 95% CI$^a$ | 95% CV$^b$ | % variance$^c$ | Q statistic |
| Between-study analyses |
| Sex of participants |
| Léger$^d$ | Men | 24 | 782 | .80 | .88 | .81-.94 | .59-1.00 | 18.35 | 124.53* |
| | Women | 13 | 475 | .74 | .81 | .70-.92 | .50-1.00 | 21.56 | 55.14* |
| Eurofit$^e$ | Men | 6 | 125 | .61 | .68 | .49-.87 | .41-0.95 | 57.29 | 5.44 |
| | Women | 4 | 98 | .67 | .75 | .53-.97 | .38-1.00 | 32.74 | 10.00* |
| Age of participants |
| Léger$^d$ | Children | 28 | 1,335 | .72 | .78 | .72-.85 | .50-1.00 | 22.30 | 113.74* |
| | Adults | 20 | 887 | .86 | .94 | .87-1.00 | .72-1.00 | 12.72 | 160.02* |
| Eurofit$^e$ | Children | 7 | 143 | .61 | .68 | .52-.84 | .50-0.86 | 76.18 | 2.66 |
| | Adults | 4 | 135 | .71 | .79 | .56-1.00 | .43-1.00 | 24.49 | 15.00* |
| Level of VO$_2$max |
| Léger$^d$ | Low | 22 | 1,181 | .75 | .82 | .74-.89 | .49-1.00 | 13.27 | 167.60* |
| | High | 23 | 895 | .80 | .88 | .80-.95 | .62-1.00 | 18.84 | 115.48* |
| Eurofit$^e$ | Low | 5 | 108 | .69 | .77 | .59-.94 | .41-1.00 | 36.29 | 10.68* |
| | High | 6 | 170 | .64 | .71 | .52-.91 | .45-0.98 | 48.32 | 7.81 |
| Within-study analysis |
| Number of predictors$^l$ |
| Léger$^d$ | One | 8 | 742 | .70 | .77 | .68-.86 | .51-1.00 | 15.56 | 50.61* |
| | Few | 8 | 742 | .80 | .88 | .80-.95 | .67-1.00 | 12.60 | 64.68* |

Note: K, number of rs; N, total sample size; $r_c$, overall weighted mean of $r$ corrected for sampling error only; $r_p$, overall weighted mean of $r$ corrected for sampling error and measurement error of the 20-m shuttle run test; 95% CI, 95% confidence interval; 95% CV, 95% credibility interval. Percentage of variance accounted for by statistical artefacts including sampling error and measurement error of the 20-m shuttle run test; *Percentage of variance accounted for by statistical artefacts including sampling error and measurement error of the 20-m shuttle run test; †Léger’s protocol starts at 8.5 km/h and increases 0.5 km/h each minute (Léger et al., 1984, 1988); Eurofit protocol starts at 8.0 km/h and increases 0.5 km/h each minute, but the second stage increases by 1.0 km/h (Council of Europe Committee for the Development of Sport, 1988); Performance only score (“one”) or performance score plus other (“few”). †Because some studies mixed categories or some values were missing, the overall n for some categories is lower for some 20-m shuttle run tests. * p < 0.05

the level of VO$_2$max (i.e. low average level, < P$_{50}$ and high average level, > P$_{50}$) on overall criterion-related validity correlation coefficients for estimating VO$_2$max for each 20MSR protocol potentially affected by moderator effects (i.e. Léger’s and Eurofit protocols). Additionally, for Léger’s protocol the correlation coefficients of criterion-related validity which were reported for both the performance score only and multiple predictors were compared (i.e. within-study analysis).

Sex of participants: The results showed that the analyzed 20MSR protocols had a moderate-to-high mean correlation coefficient of criterion-related validity for estimating VO$_2$max for both men and women in which all 95% CI did not include zero. Moreover, all the 95% CI of mean correlation coefficients overlapped. On the other hand, we must point out that, according to moderator analysis criteria, at least two of the three criteria were met in the 20MSR test, indicating that the criterion-related validity of these protocols separately for sex was still heterogeneous. Finally, because some studies grouped men and women together or the sex values were missing, in Table 2 overall n for sex of participants is lower.

Age of participants: Results showed that Léger’s protocol had a moderate mean correlation coefficient of criterion-related validity for estimating VO$_2$max for children and moderate-to-high for adults. Additionally, the results of the present meta-analysis showed that the criterion-related validity of Léger’s protocol was statistically higher for adults than for children. On the contrary, the results showed that the Eurofit protocol had a moderate mean correlation coefficient of criterion-related validity for estimating VO$_2$max for both children and adults, as
well as the 95% CI of mean correlation coefficients overlapped. In addition, the 95% CIs included the value zero. Finally, according to moderator analysis criteria, at least one of the three criteria was met in all categories, indicating that the criterion-related validity of these protocols separately for age was still heterogeneous.

**Level of VO_{2max}:** The results showed that the analyzed 20MSR protocols had a moderate-to-high mean correlation coefficient of criterion-related validity for both participants with low and high level of VO_{2max} in which all 95% CI did not include the value zero. Furthermore, all the 95% CI of mean correlation coefficients overlapped. Regarding the moderator analysis criteria, at least two of the three criteria were met in all categories, indicating that the criterion-related validity of these 20MSR protocols separately for level of VO_{2max} were still heterogeneous. Finally, because for Léger’s protocol the 95% CI overlapped, the large CI probably because of the low number of correlations found must be highlighted. Additionally, any primary study comparing the criterion-related validity of various protocols of the 20MSR test was found. Therefore, we should be cautious with the overall results of the present meta-analysis.

**Number of predictors:** The results showed that for Léger’s protocol the performance score only had a moderate mean correlation coefficient of criterion-related validity for estimating VO_{2max} mainly when, the performance score was added to other variables the criterion-related validity was moderate-to-high. However, the 95% CI of mean correlation coefficients overlapped. Additionally, the three criteria were met in both categories, indicating that the criterion-related validity was heterogeneous.

In summary, the overall results showed that the 20MSR test had a statistically significant and moderate-to-high criterion-related validity for estimating VO_{2max}. The choice of a cardiorespiratory fitness test must be based on its functionality and validity. Although the VO_{2max} measured during a laboratory-based and graded maximal exercise test has the advantage of being the criterion measure to assess cardiorespiratory fitness, due to several practical reasons they have the disadvantage of having a limited use in several settings (Pescatello et al., 2014). In settings such as sports clubs, schools or large scale research studies, as the 20MSR test has the advantage of allowing for an evaluation in a short amount of time with minimal skill and instrumentation, potentially it could be a useful alternative to estimate cardiorespiratory fitness. In this context, the overall results of the present meta-analysis show that the 20MSR test has a moderate-to-high mean correlation coefficient of criterion-related validity for estimating VO_{2max}.

Since the original 20MSR test of one-min stages (Léger et al., 1984), various modifications of the start and subsequent speed increases have been proposed (e.g. Council of Europe Committee for the Development of Sport, 1988; Dong-Ho et al., 2014; Riddoch, 1990). However, according to the results of the present meta-analysis (and despite the fact that we are aware that the 95% CI of mean correlation coefficients with the Eurofit protocol overlapped), Léger’s protocol showed a greater average criterion-related validity coefficient. Therefore, if our purpose is to assess cardiorespiratory fitness, it seems that the use of the Eurofit and QUB’s protocols is not justified. However, the fact that in the present meta-analysis the overall criterion-related validity of Léger’s and Eurofit protocols was heterogeneous, as well as the number of correlations found for the QUB’s protocol was low, must be highlighted. Additionally, any primary study comparing the criterion-related validity of various protocols of the 20MSR test was found. Therefore, we should be cautious with the overall results of the present meta-analysis.

The second purpose of this meta-analysis was to examine the influence of some potential moderator factors (sex, age, and level of VO_{2max} of the participants) in criterion-related validity coefficients of the 20MSR test. One of the main findings of the present meta-analysis showed that the criteria-related validity of Léger’s protocol was significantly higher for adults compared with children. Similarly, for the Eurofit protocol the average correlation coefficient for adults was considerably higher than for children. Although we have to be aware that for that protocol the 95% CI overlapped, the large CI probably because of the low number of correlations found must also be considered. Additionally, while 51% of the total simple correlation coefficients with children were equal or below 0.70 (i.e. less than 50% of variance explained), only 20% was found for adults. Therefore, the results of the present meta-analyses show that the criterion-related validity of the 20MSR test is statistically higher for adults versus children.

In line with the findings of the present meta-analysis, Matsuzaka et al. (2004) found out that, when participants were examined under the same experimental conditions (e.g. field and laboratory tests protocols, equipment, and testers), the criterion-related validity of the 20MSR test was considerably higher for adults ($r = 0.92$) than for children ($r = 0.75$) and adolescents ($r = 0.76$). Similarly, Léger et al. (1988) suggested that, since the chronological age of children, and not adults, was a significant predictor of VO_{2max}, the lower validity of the 20MSR test in children as compared to adults might be the result of larger interindividual variations. In addition to chronological age, in the present meta-analysis, it has been found that among children other variables such as sex, biological maturation, body mass, body mass index, body fat and/or skinfolds were significant predictors of the VO_{2max}. Furthermore, since children might be less willing to endure discomfort of strenuous effort, have less motivation, and/or a limited attention span for monotonous tasks, the 20MSR test performance could be affected and, therefore, its criterion-related validity.
Another potential reason for these results could be that the starting speed of the 20MSR test is too high for children. Current evidence suggests that to elicit valid VO\textsubscript{2max} values, continuous incremental tests should last at least five minutes (Midgley et al., 2008). However, Castro-Piñero et al. (2011) in a population-based study carried out using Léger’s protocol (i.e. starting speed 8.5 km/h) found that most 6- to 17-year-old children did not complete five stages (i.e. five minutes). Previous studies have proposed modifications of the 20MSR test for children with a drastically reduced starting speed (e.g. 4 km/h, Quinart et al., 2014; 6.5 km/h, Cadenas-Sánchez et al., 2014). Unfortunately, these authors did not either examine the criterion-related validity of the test (Cadenas-Sánchez et al., 2014) or did not compare it with “traditional” protocols such as the Léger protocol (Quinart et al., 2014). As regards the moderator analyses for sex and VO\textsubscript{2max} levels, according to the results of the present meta-analysis they seem not to affect the criterion-related validity. Therefore, since the criterion-related validity for both men-women and low-high level of VO\textsubscript{2max} subgroups was similar, the 20MSR test can be used interchangeably for any subcategory. Regarding the VO\textsubscript{2max} categories, however, due to the fact that in the present meta-analysis the n was classified based on the average score, we have to be aware that several participants with low VO\textsubscript{2max} values could be classified as high values and vice versa. This fact could affect the results of the present meta-analysis.

Finally, the third purpose of the present meta-analysis was to compare the values of the criterion-related validity coefficients between the performance only score and the performance score combined with other variables. When multiple predictors were used, the average correlation coefficient was considerably higher than for the performance score only (on average r\Delta = 0.11). Although we have to be aware that for that protocol the 95% CI overlapped, the large CI, probably because of the low number of correlations found must also be considered. It must be also pointed out that seven of the eight studies analyzed were carried out with children. Thus, children seem to benefit considerably from other variables to estimate cardiorespiratory fitness. In summary, although the criterion-related validity of the 20MSR test is statistically lower for children than for adults, when the performance score is combined with other variables such as age, sex, body mass or body mass index, the criterion-related validity value of the 20MSR test is considerably high.

**Strengths and limitations**

The meta-analysis is a useful tool to assess scientific evidence, but an understanding of its strengths and limitations is needed for the most appropriate use of this method. An extensive revision of the general strengths and limitations of meta-analyses (e.g. Cooper et al., 2009), as well as specifically in the meta-analysis of the criterion-related validity of physical fitness field tests has been published elsewhere (Mayorga-Vega et al., 2014a).

Briefly, regarding the strengths of the present meta-analysis, we followed several measures to avoid (or at least to reduce) publication bias. Firstly, to avoid availability bias, we conducted a wide literature search through several databases without limiting any kind of manuscript, language or publication date. Due to the limitations of databases to find the “fugitive” literature, several complementary searches were also carried out. Secondly, in the present meta-analysis all the studies published by the same authors were thoroughly cross-referenced with each other in order to avoid duplication. Lastly, several exploratory analyses were also conducted to identify and assess the impact of any potential publication bias.

Another strength of the present meta-analysis is related to the statistical approach used. In the present study, the Hunter-Schmidt’s psychometric meta-analysis approach (2004) was conducted in order to obtain the population estimates of criterion-related validity of the 20MSR test. Since this method estimates the population correlation by correcting the observed correlations due to various artefacts such as sampling error and measurement error, it has been considered one of the best meta-analyses approaches.

Regarding the limitations, there were some that should be considered when examining the results of the present meta-analysis. The main limitations were related to the small number of criterion-related validity coefficients found. Estimating the population parameters based on small samples is simply less accurate than in a large-sized meta-analysis. Because a partial hierarchical breakdown (instead of full) had to be used, misleading results due to confounding and interaction effects might be produced (Hunter and Schmidt, 2004). Therefore, the results of the present study should be considered with caution; firmer conclusions should await the accumulation of a larger number of studies (Hunter and Schmidt, 2004).

Another limitation of the present meta-analysis is related to the criterion measure used in the studies. Although in the present meta-analysis only primary studies that used as the criterion measure, the VO\textsubscript{2max} during a standardized and laboratory-based incremental test to exhaustion were selected (see results section), in these studies different equipment (various brand and characteristics), ergometers (i.e. treadmill and cycle ergometer) and protocols (e.g. in warm-up, initial load, increasing load, gas collection time, or number of gas collections) were used. Furthermore, in the studies there is not a wide agreement about the criteria to determine VO\textsubscript{2max}. For example, researchers used a plateau in VO\textsubscript{2}, the respiratory exchange ratio, or the age-adjusted estimates of the maximal heart rate, alone or in combination; then, the quantitative cut-off values criteria are also diverse.

The fact that in the present meta-analysis peak oxygen uptake (VO\textsubscript{peak}) has been used interchangeably with VO\textsubscript{2max} must be highlighted. Although we are aware that the VO\textsubscript{peak} simply refers to the highest value of oxygen uptake (VO\textsubscript{2}) attained on a particular exercise test, due to the fact that the tests in the primary studies of the present meta-analysis were maximal we can be reasonably sure that values were the highest value of VO\textsubscript{2} that is deemed attainable by individuals, i.e. the VO\textsubscript{2max} (Rowland, 1993). Therefore, it seems that the criterion measure of cardiorespiratory fitness should be reexamined and readjusted (Howley et al., 1995).
Finally, coding some study features was problematic due to different reasons. For instance, because in the present meta-analysis the level of VO2max was classified based on the average scores, we are aware that several individuals with low VO2max could be classified as high VO2max and vice versa. Additionally, although participant characteristics such as physical activity levels or sport participation were potentially moderating features, coding for them was not possible because most studies did not identify them.

Conclusion

Overall the 20MSR test has a moderate-to-high mean correlation coefficient of criterion-related validity for estimating VO2max. Regarding the potential moderators examined, the present meta-analysis shows that the criterion-related validity of the 20MSR test is higher for adults than for children. Nevertheless, when the performance score among children is combined with other variables, the criterion-related validity to estimate the VO2max is considerably high. As regards the sex and level of VO2max of participants, they seem not to affect the relationship between the 20MSR test score and the measured VO2max.

When an individuals’ VO2max attained during a laboratory maximal exercise test is not feasible such as in sports clubs, schools or large scale research studies, scientists and practitioners could use the 20MSR test as a useful alternative to estimate cardiorespiratory fitness. Among adults the performance only score seems to be a strong estimator of cardiorespiratory fitness, in contrast among children the performance score should be combined with other variables. Nevertheless, as in the application of any physical fitness field test, testers must be aware that the performance score of the 20MSR test is simply estimation and not a direct measure of cardiorespiratory fitness.

Due to the relatively low number of r values found and that criterion-related validity of the 20MSR test with in most categories is heterogeneous, we should be cautious with the results of the present meta-analysis. Therefore, when a greater number of studies are accumulated, a large sized meta-analysis with a full hierarchical analysis approach should be carried out. For this purpose future research studies should further examine the criterion-related validity of the 20MSR test, especially in modifications of the test with a lower starting speed, among populations such as children, and go deeper into other related aspects such as the potential moderator effects of the level of VO2max.

Acknowledgements

We gratefully acknowledge all the authors of the original research studies for their contribution, without whom the present meta-analysis could not be done. We especially acknowledge Professor Luc Léger for his gracious help in the identification of several synonyms for the test. Furthermore, we thank the head of the library, Mrs. Ana M. Peregín González, for technical assistance and her great help to retrieve manuscripts that were not readily available. Additionally, we would like to express our gratitude to the Associate Editor-in-Chief Roger Ramsbottom for all the hard work and detailed care that have undoubtedly improved the present manuscript. We also thank Alissa Hatten-Viciana for the English revision. Daniel Mayorga-Vega is supported by the Spanish Ministry of Education, Culture and Sport (AP2010-5905).

References


tal exercise test for valid VO_{2max} determination really need to last between 8 and 12 minutes? Sports Medicine 38, 441-447.


