



UNIVERSITY OF
GLOUCESTERSHIRE

This is a peer-reviewed, post-print (final draft post-refereeing) version of the following published document and is licensed under Creative Commons: Attribution 4.0 license:

Martinez-Romero, Maria T, Ayala, Francisco, De Ste Croix, Mark B ORCID: 0000-0001-9911-4355, Vera-García, Francisco J, Sainz de Baranda, Pilar, Santonja-Medina, Fernando and Sanchez-Meca, Julia (2020) A meta-analysis of the reliability of four field-based trunk extension endurance tests. International Journal of Environmental Research and Public Health, 17 (9). E3088. doi:10.3390/ijerph17093088

Official URL: <https://doi.org/10.3390/ijerph17093088>

DOI: <http://dx.doi.org/10.3390/ijerph17093088>

EPrint URI: <https://eprints.glos.ac.uk/id/eprint/8322>

Disclaimer

The University of Gloucestershire has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

The University of Gloucestershire makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

The University of Gloucestershire makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

The University of Gloucestershire accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

PLEASE SCROLL DOWN FOR TEXT.

A Meta-Analysis of the Reliability of Four Field-Based Trunk Extension Endurance Tests

María Teresa Martínez-Romero ¹, Francisco Ayala ^{2,*}, Mark De Ste Croix ³, Francisco José Vera-García ⁴, Pilar Sainz de Baranda ⁵, Fernando Santonja-Medina ⁶ and Julio Sánchez-Meca ⁷

- ¹ Department of Physical Activity and Sport, Faculty of Sports Sciences, University of Murcia, 30720 San Javier (Murcia), Spain; mariateresa.martinez13@um.es
- ² Department of Sport Science, Sports Research Centre, Miguel Hernández University of Elche, 03202 Elche (Alicante), Spain; franciscoayalarodriguez@gmail.com
- ³ School of Sport and Exercise, Exercise and Sport Research Centre, University of Gloucestershire, GL2 9HW Gloucester, United Kingdom; mdestecroix@glos.ac.uk
- ⁴ Department of Sport Science, Sports Research Centre, Miguel Hernández University of Elche, 03202 Elche (Alicante), Spain; fvera@goumh.umh.es
- ⁵ Department of Physical Activity and Sport, Faculty of Sports Sciences, University of Murcia, 30720 San Javier (Murcia), Spain; psainzdebaranda@um.es
- ⁶ Department of Surgery, Pediatrics, Obstetrics and Gynecology, Faculty of Medicine, University of Murcia, 30100 Murcia, Spain; fernando@santonjatrauma.es
- ⁷ Department of Basic Psychology and Methodology, Faculty of Psychology, University of Murcia, 30100 Murcia, Spain; jsmecca@um.es
- * Correspondence: franciscoayalarodriguez@gmail.com

Received: 15 April 2020; Accepted: 26 April 2020; Published: date

Abstract: This meta-analysis aimed to estimate the inter- and intra-tester reliability of endurance measures obtained through trunk extension field-based tests and to explore the influence of the moderators on the reliability estimates. The reliability induction rate of trunk extension endurance measures was also calculated. A systematic search was conducted using various databases, and subsequently 28 studies were selected that reported intraclass correlation coefficients for trunk extension endurance measures. Separate meta-analyses were conducted using a random-effects model. When possible, analyses of potential moderator variables were carried out. The inter-tester average reliability of the endurance measure obtained from the Biering-Sorensen test was intraclass correlation coefficient (ICC) = 0.94. The intra-session reliability estimates of the endurance measures recorded using the Biering-Sorensen test, the prone isometric chest raise test, and the prone double straight-leg test were ICC = 0.88, 0.90, and 0.86, respectively. The inter-session average reliability of the endurance measures from the Biering-Sorensen test, the prone isometric chest raise test, and the dynamic extensor endurance test were ICC = 0.88, 0.95, and 0.99, respectively. However, due to the limited evidence available, the reliability estimates of the measures obtained through the prone isometric chest raise, prone double straight-leg, and dynamic extensor endurance tests should be considered with a degree of caution. Position control instruments, tools, and familiarization session demonstrated a statistical association with the inter-session reliability of the Biering-Sorensen test. The reliability induction rate was 72.8%. Only the trunk extension endurance measure obtained through the Biering-Sorensen test presented sufficient scientific evidence in terms of reliability to justify its use for research and practical purposes.

Keywords: reliability generalization meta-analysis; core endurance; physical education; sports performance; pre-participation assessment

Commented [M1]: Please carefully check the accuracy of names and affiliations.

1. Introduction

In the last two decades, there has been an increased interest in the assessment of trunk extensor muscle endurance, as deficits in trunk extensor endurance and imbalances between trunk muscle groups have been suggested as being primary risk factors for low back disorders [1–4] and could negatively affect sports performance [5,6].

Although several sophisticated laboratory-based tests have been developed to quantify trunk extensor muscle endurance (i.e., force platforms, isokinetic dynamometers) [7–9], field-based tests seem to be the most popular tests, probably because they are portable, cost-effective, easy to use, and time-efficient methods [10]. Different trunk extension endurance field-based tests have been widely used, including (a) isometric endurance tests (i.e., Biering-Sorensen test [2], prone isometric chest raise test [11], and prone double straight-leg raise test [12]), which involve maintaining a position against gravity for as long as possible, and (b) dynamic endurance tests (i.e., dynamic extensor endurance test [3,13]), which consist of performing as many repetitions as possible in a given time or with a certain cadence until exhaustion. However, the validity and reliability of their outcomes must be determined [14] before these field-based tests can be used to identify deficits in trunk extensor muscle endurance, in order to explore imbalances between trunk muscle groups and to establish progress from training and/or rehabilitation programs. The four trunk extension endurance measures identified have been considered operationally valid by medical (ACSM [15]), sport (Swiss Olympic Medical Centers [16]), and educational (Cooper Institute [17]) organizations. Furthermore, their measures have been shown to be sensitive enough to detect trunk extensor muscle endurance deficits in patients with chronic low back pain [18,19], and are included in many prominent sports medicine textbooks [15–17].

Two different types of reliability should be considered before choosing an appropriate test for research and clinical purposes [20,21]: inter-tester reliability and intra-tester reliability. Inter-tester reliability provides information regarding the degree to which measures taken by different testers using identical test protocols on the same cohort of individuals are similar or consistent [20]. Intra-tester reliability provides information regarding the degree to which several measurements taken at different times for the same test by the same tester are similar [22]. The intra-tester reliability can be determined using short (generally within a day: intra-session) or long (generally more than one day: inter-session or test–retest) time intervals to separate the testing sessions [14,20].

In the scientific literature, several studies have examined the inter- [18,23,24] and intra-tester (intra- [18,25,26] and inter-session [9,11,27]) reliability of the four field-based trunk extension endurance tests in different cohorts of individuals (e.g., athletes, children, adolescents, sedentary adults). The findings of these studies, and for both types of reliability, have shown a large degree of heterogeneity, with intraclass correlation coefficient (ICC) scores ranging from 0.77 to 0.99 and from 0.20 to 0.99 for the inter- and intra-tester reliability, respectively.

The large degree of heterogeneity presented in the ICC scores reported suggests that the reliability outcomes of a physical performance measure could be influenced by several factors (the tester's experience or training in administering the test, variations in the assessment methodology, and participant-related variability, etc.). If practitioners do not consider these factors, they might select a trunk extension endurance test with an inter- and intra-tester reliability that is inappropriate for their given population. In practical terms, the selection of an inappropriate test can lead not only to an inaccurate diagnosis of normality or deficits but also an incorrect evaluation of the effectiveness of a training or rehabilitation intervention for improving or maintaining trunk extensor endurance. Frequently, studies exploring trunk extension endurance do not report the reliability estimates of the test they are using with the given population [28–30], or the reliability coefficients obtained in previous studies are simply cited [3,31,32]. This characteristic of referencing the reliability coefficients from prior studies instead of reporting the reliability obtained with their data has been coined reliability induction (RI) [33].

The reliability generalization (RG) is a meta-analytical approach that emerges as a criticism of the widespread practice of RI. The purpose of this method is to estimate the average reliability of the scores of a given test, as well as to determine the variability of the reliability coefficients reported by

Commented [AT2]: Please check the intended meaning is retained.

the different studies that have used this test. Moreover, if the variability is very high, another aim is to explore which characteristics of the studies may be statistically associated with the reliability estimates [34–36]. To the best of the authors' knowledge, there are no RG meta-analyses published to date concerning trunk extension endurance measures obtained through field-based tests. This information might be useful for practitioners because it can aid them in deciding on the best field-based test/s (in term of reliability) to assess trunk extensor muscle endurance according to the characteristics of their patients or athletes.

Therefore, the main purpose of the current study was to conduct a RG meta-analysis (a) to obtain combined reliability estimates of trunk extension endurance measures obtained through four field-based tests, (b) to identify which characteristics of the studies may influence the variability of the reliability coefficients, and (c) to determine the RI practice in studies that have used trunk extensor muscle endurance tests.

2. Materials and Methods

The current RG meta-analysis (PROSPERO ID: CRD42019123179) was conducted following the PRISMA guidelines [37]. The PRISMA checklist is presented in Table S1 in the Supplementary Materials section.

2.1. Study Selection

In this RG meta-analysis, the following inclusion criteria were considered: (a) being an empirical research study (psychometric study) in which the original and modified versions of the following field-based tests were applied: Biering-Sorensen test [2], prone isometric chest raise test [11], prone double straight-leg raise test [12], and dynamic extensor endurance test [3] (a detailed description of each test is presented in **Box 1**); (b) being written in English or Spanish; (c) being published before March 2019; (d) using a sample of at least 10 participants; and (e) reporting the ICC as a measure of reliability (for both the inter- and intra-tester (intra-session or internal consistency and inter-session or test-retest reliability) reliability) of any of the aforementioned field-based tests or providing sufficient data from which this coefficient could be calculated through standardized equations. The selection of the ICC as a valid measure of reliability was made following the consensus-based standards for the selection of health measurement instruments (COSMIN) statement [38–40]. Studies whose main objective was not the analysis of the reliability of the trunk extension endurance measures obtained through any of the four aforementioned field-based tests, but which reported reliability data were also eligible for inclusion. The same inclusion criteria were considered for selecting studies that induced reliability, except (d) and (e).

Literature reviews, abstracts, editorial commentaries, and letters to the editor were excluded. Finally, some authors were contacted to provide missing data or to clarify if data were duplicated in other publications. Incomplete data or data from an already included study were also excluded.





2.2. Search Strategy

A systematic computerized search was conducted up to 28th February 2019 in the databases MEDLINE, PubMed, Web of Science, and Scopus. In addition, a supplementary search in Google Scholar was also performed. Relevant keywords were used to construct Boolean search strategies, including terms such as reliability, extensor muscles, trunk, core, endurance, performance, and test (see Table S2). Then, the reference lists of included studies were screened for potentially eligible studies. Two reviewers independently (M.T.M. and P.S.B.) selected studies for inclusion in a two-step process. First, studies were screened on the basis of title and abstract. In a second stage, full-text studies were reviewed to identify those studies that met the eligibility criteria. Disagreements were resolved through consensus or by consulting a third reviewer (F.A.).

Box 1. Trunk extension endurance field-based tests description.

Isometric Endurance Field-Based Tests

Commented [M3]: Due to journal' regulations, we changed "table 1" to "box 1", please confirm.

	<p>Biering-Sorensen test</p> <p>The test consists of assessing how many seconds the participant can keep the unsupported upper part of the body (from the upper border of the iliac crest) horizontal while placed prone with the buttocks and legs fixed to the table bench by three wide canvas straps, with the arms across the chest. The test is continued until the participant could no longer control his/her posture for a maximum of 240 s.</p>
	<p>Prone isometric chest raise test</p> <p>The test consists of assessing how many seconds the participant can keep the sternum off the floor while placed prone with the arms along the body. A small pillow is placed under the iliac crest to decrease the lumbar lordosis. The subject is asked to maintain the position for as long as possible, not exceeding a 5 min time limit.</p>
	<p>Prone double straight-leg raise test</p> <p>The test consists of assessing how many seconds the participant can keep both legs raised with the knees off the mat while placed prone with hips extended, the hands underneath the forehead and the arms perpendicular to the body. The test is continued until the participant can no longer maintain knee clearance.</p>
Dynamic Endurance Field-Based Test	
	<p>Dynamic extensor endurance test</p> <p>The test consists of assessing how many repetitions the participant can perform while placed prone with the unsupported upper part of the body (from the upper border of the iliac crest). The arms are positioned along the body and the buttocks and legs are fixed by three straps. With the spine kept straight, the subject is instructed to extend the trunk to neutral and then to lower the upper body 45 degrees. A repeated beat guided the subject to maintain a cadence of 25 repetitions per minute until exhaustion.</p>

2.3. Data Extraction

To guarantee the maximum possible objectivity, a codebook was produced that specified the standards followed in coding each of the characteristics of the studies selected (see [Box S1](#)). The moderator variables of the eligible studies were coded and grouped into four categories: (1) general study descriptors (authors, publication year, country, and study objective), (2) description of the study population (sample size, age, sex distribution, and target population), (3) description of the field-based test (version, protocol, and scores), and (4) type of reliability analyzed (inter-tester and/or intra-tester) and main characteristics of the study design (familiarization sessions, number of measurements, number of testers, time interval between measurements, test conditions, etc.).

2.4. Reliability Estimates

In this RG study, the ICC was extracted from the eligible studies as the only valid measure of inter- and intra-tester reliability. There are other coefficients such as the standard error of measure (SEM), Pearson correlation, and minimal detectable change with a 95% confidence level (MDC_{95}) that

Commented [M4]: We changed "Table S3" to "Box S1", please confirm.

have also been proposed as measures of reliability [20,41]. However, the scarce number of references that reported these types of reliability coefficients [42–44] did not allow for any separate meta-analyses to be conducted.

In order to maximize the number of studies included in this RG, when an article only reported the SEM, ICC was estimated by

$$ICC = \sqrt{1 - \frac{SEM^2}{SD^2}}, \quad (1)$$

with *SD* being the standard deviation of the test scores.

Likewise, when an article only reported the MDC, it was transformed into SEM using the following formula:

$$SEM = \frac{MDC}{1.96 \times \sqrt{2}} \quad (2)$$

Later, the SEM was converted into ICC using the previously stated formula.

2.5. Quality Assessment

The COSMIN methodology was used to evaluate the quality of eligible studies for the RG meta-analysis, which consists of three sub-steps [38–40]. First, the risk of bias in each study was assessed using the COSMIN risk of bias check-list [39]. The check-list contains standards referring to design requirements and preferred statistical methods of studies on measurement properties. For each measurement property, a COSMIN box was developed, containing all standards needed to assess the quality of a study on that specific measurement property. Specifically, Box 6 regarding reliability was used for this RG meta-analysis. Each standard of the box was rated as very good, adequate, doubtful, or inadequate quality. To determine the overall rating of the quality of every single study, the lowest rating of any standard in the box was taken (i.e., “the worst score counts” principle) [45]. Second, from the data extracted on the description of the study population and the results on the reliability coefficients, the result per study was rated against the criteria for good measurement properties [38] as sufficient (+ (ICC ≥ 0.70)), insufficient (− (ICC < 0.70)), or indeterminate (? (ICC not reported)). Finally, the results from different studies on one measurement property were statistically pooled in a meta-analysis and the quality of the evidence was graded (high, moderate, low, or very low evidence) using a modified GRADE approach. Four of the five GRADE factors were used in this meta-analysis: risk of bias (i.e., the methodological quality of the studies), inconsistency (i.e., unexplained inconsistency of results across studies), imprecision (i.e., total sample size of the available studies), and indirectness (i.e., evidence from different populations than the population of interest in the review). The fifth factor, publication bias, is difficult to assess in studies on measurement properties due to a lack of registries for these types of studies. Therefore, we did not take this factor into account in this meta-analysis [38–40]. The starting point is always the assumption that the pooled or overall result is of high quality. The quality of evidence is subsequently downgraded by one or two levels per factor to moderate, low, or very low when there is a risk of bias, inconsistency, imprecision, or indirect results [38]. Table S3 displays a brief description of each step of the COSMIN methodology.

The data extraction and quality assessment were double-coded (M.T.M. and P.S.B.) to assess the inter-coder reliability of the coding process. Two authors working independently randomly coded 50% of the studies. For the quantitative moderator variables, ICC_{3,1} were calculated, whereas for the qualitative moderator variables, we applied Cohen’s kappa coefficients. In general, the agreement coefficients between the two authors were satisfactory as proposed by Orwin and Vevea [46], with the kappa coefficients ranging between 0.74 and 1, and the ICC ranging between 0.96 and 1. Inconsistencies between the two coders were resolved by consensus, and when these were due to ambiguity in the coding book, this was corrected. As before, any disagreement was resolved by mutual consent or in consultation with a third reviewer (FA).

2.6. Data Synthesis and Analysis

Commented [M5]: Please write the equation with formula editor, like MathType.

Commented [AT6]: Please define as appropriate.

Separate meta-analyses were carried out for the different field-based tests and types of reliability analysis to avoid dependence problems, given that each study could report more than one reliability coefficient.

Statistical analyses were completed by assuming a random-effects model, and each reliability coefficient was weighted by the inverse variance [47]. Reliability coefficients were transformed into Fisher's Z to normalize their distribution and to stabilize the variances [36]. For each meta-analysis, an average reliability coefficient (ICC+) and a 95% confidence interval (95% CI) were calculated [48].

The heterogeneity exhibited by the reliability coefficient, which represents the percentage of total variation across all studies due to between-study heterogeneity, was assessed by constructing a forest plot and by calculating the Q statistic and the I^2 index. Analyses of potential moderator variables were carried out when the Q statistic was statistically significant, the I^2 index was over 25%, and there were at least 20 reliability coefficients [49]. The only test that fulfilled these conditions was the Biering-Sorensen test for intra-tester reliability ($k = 28$ studies).

The influence of qualitative moderator variables (i.e., sample type (children and adolescents, adults), target population (community, clinical), physical activity level (sedentary, active), test duration (until exhaustion, until 240 s), among others) on the reliability coefficients were explored through analysis of variance (ANOVAs) and assuming mixed-effects model. Meta-regressions were also applied to test the influence of continuous moderators on the reliability coefficients assuming mixed-effects model, such as final sample size, average percentage of females, average age, percentage of attrition, number of measurements, number of testers, and time interval between measurements. Q_B and Q_w for ANOVAs and Q_R and Q_E statistics for meta-regressions were calculated to test the statistical significance of each moderator variable and to assess the model misspecification, respectively [50]. In addition, the proportion of variance accounted for by the moderator variables was estimated with R^2 following Raudenbush's proposal [51]. In order to find the subset of moderator variables that can explain most of the reliability coefficient variability, a multiple meta-regression model (by assuming a mixed-effects model) was adjusted.

To facilitate the interpretation of the results, the average reliability coefficients and their confidence limits were back-transformed to the original metric of reliability coefficients. Furthermore, it has been suggested that for studies conducted in the sports medicine field of knowledge and aimed at analyzing the reliability (inter- and intra-tester) of quantitative physical performance measures, ICC values of 0.8 to 0.9 may be considered as acceptable, but values higher than 0.9 are desirable [14].

The statistical analyses were carried out with the software Comprehensive Meta-analysis 3.3 [52].

Commented [M7]: Please state manufacturer, city and country from where software has been sourced.

3. Results

3.1. Study Selection

A total of 1452 references were identified with all search strategies, from which 494 were excluded in the first screening as duplicates (34%). A total of 674 studies (46.4%) were eliminated after reading the title and abstract. Another 181 studies (12.4%) were removed after reading the full-text—120 studies did not use any trunk extensor muscle endurance field-based test (8.2%) and 61 articles did not meet the established inclusion criteria (4.2%).

This search process identified 103 empirical studies that met the inclusion criteria (7.1%), in which 28 articles (1.9%) reported some reliability coefficients [8–11,18,23–27,43,53–69] (resulting in 43 cohort groups, as 9 studies had more than 1 group) and 75 articles (5.1%) induced the reliability (see the references in Document S1 in the Supplementary Materials section). Figure 1 shows the flow chart of the selection process of the studies.

Commented [AT8]: Please check the intended meaning is retained.

Commented [M9]: We changed S5 to S1, please confirm.

3.2. Descriptive Characteristics of the Selected Studies for the RG Meta-Analysis

Table S4 provides a descriptive summary of the characteristics of the included studies for the RG meta-analysis. The studies selected were carried out between 1986 and 2018 and comprised participants from four continents (Europe [8,10,24,43,53–57,59–61,63,65,66], America [9,25–27,64,67–

69], Asia [11,18,58,62], and Oceania [23]). Only one study was written in Spanish [60], and the rest of the selected studies were written in English [8–11,18,23–27,43,53–59,61–69].

The total sample size was 1097 participants, with an average of 25.5 subjects per cohort group (minimum = 10 [43,53,59,62,63] and maximum = 100 [11]). Three studies used healthy children [57] and adolescents [27,62]. Furthermore, 14 studies employed asymptomatic samples [9,10,18,24,25,27,53,56,57,59,60,62,63,68,69], 8 used patients with LBP [8,54,55,58,64–67], and 6 included both types of samples [11,23,26,43,61].

Concerning the type of field-based test used, 25 studies (89.2%) investigated the reliability of the back extensor endurance measures obtained through different modified versions of the Biering-Sorensen test [8–10,18,23–27,43,53,54,56,57,59–69], 4 studies (14.3%) used the prone isometric chest raise test (two of them used the original version [11,58] and another two used a modified version [18,55]), 2 studies employed a modified version of the dynamic extensor endurance test [27,62], (7.1%) and 1 study used the original prone double straight-leg raise test [18] (3.6%).

The most common types of reliability coefficients reported in the selected articles were the intra-tester and inter-session reliability (24 studies [8–11,24,26,27,43,53–66,68,69] that reported 37 coefficients), followed by the intra-tester and intra-session reliability (5 studies [18,25,26,43,67] that reported 16 coefficients), and the inter-tester reliability with 10 coefficients reported from 5 different studies [18,23,24,26,66].

3.3. Quality of the Selected Studies for the RG Meta-Analysis

Regarding the use of the Biering-Sorensen test in the studies, the COSMIN risk of bias checklist reported very good methodological quality scores for seven studies (five studies that evaluated the intra-tester and the inter-session reliability [8,10,59,61,66] and two studies for the inter-tester reliability [24,66]) and adequate quality scores in one study that analyzed the intra-tester and inter-session reliability [26]. In contrast, 13 studies [9,24,27,43,53,54,56,57,60,64,65,68,69] that analyzed the intra-tester and inter-session reliability obtained a doubtful methodological quality, as well as all of the studies that assessed the intra-tester and intra-session reliability [18,25,26,43,67] and three studies that evaluated the inter-tester reliability [18,23,26]. Two studies that analyzed the intra-tester and inter-session reliability showed inadequate methodological quality [62,63].

Commented [AT10]: Please define as appropriate.

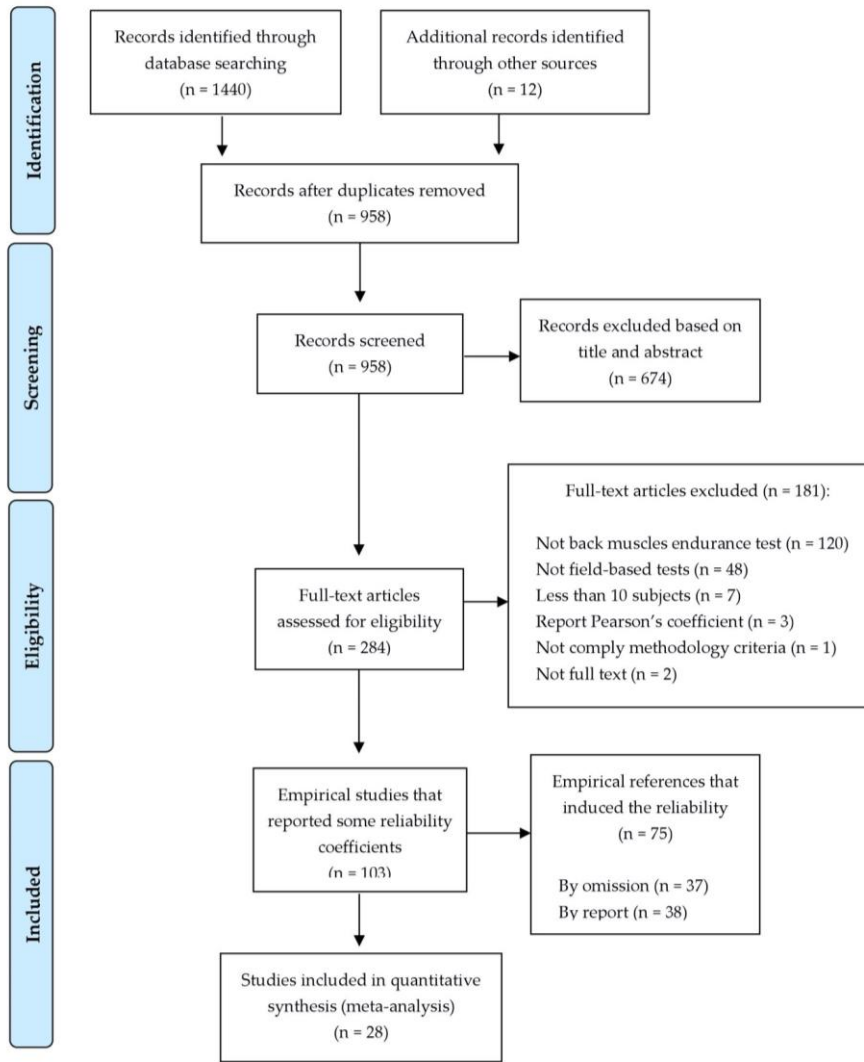


Figure 1. PRISMA flow diagram of the literature search.

Concerning the use of the prone isometric chest raise test, three studies presented doubtful methodological quality scores for intra-tester and inter-session reliability [11,55,58], the same as for intra-tester and intra-session and inter-tester reliability where only one study was found [18]. It was the same with the prone double straight-leg test, which reported a doubtful quality for the only study found [18]. Finally, the two studies that used the dynamic extensor endurance test exhibited doubtful [27] and inadequate [62] methodological qualities for the intra-tester and inter-session reliability.

The detailed data of the COSMIN risk of bias check-list and criteria for good measurement properties are presented in Table S5. Likewise, a summary of findings (SoF) per type of reliability and field-based test, including the pooled results of the measurement properties, the overall rating

Commented [AT11]: Please change “Records excluded based on” to “Records excluded on the basis of” in the figure.

Commented [AT12]: Please define as appropriate.

based on the inconsistency of results (i.e., sufficient (+), insufficient (-), or indeterminate (?)), and the grading of the quality of evidence (i.e., high, moderate, low, very low) is presented in Table 1. In general, the Biering-Sorensen test is the only test that presented moderate quality of evidence for inter-tester and intra-tester (inter-session) reliability.

Table 1. Summary of findings (SoF).

Reliability	Pooled Result ^a	Overall Rating ^b	Quality of Evidence ^c
Biering-Sorensen Test			
Inter-tester reliability	ICC = 0.94 (0.84–0.98) Consistent results Sample size = 215	Sufficient	Moderate (as there are multiple doubtful [18,23,27] and two very good studies [64,69])
Intra-tester (intra-session) reliability	ICC = 0.88 (0.83–0.92) Consistent results Sample size = 258	Sufficient	Low (as all studies are doubtful [18,25,27,44,70])
Intra-tester (inter-session) reliability	ICC = 0.88 (0.80–0.92) Consistent results Sample size = 688	Sufficient	Moderate (as there are multiple doubtful and five very good studies [8,10,61,63,69])
Prone Isometric Chest Raise Test			
Inter-tester reliability	ICC = 0.90 (0.80–0.95) Sample size = 30	Indeterminate	Did not pool the results or grade the evidence due to there being one study available [18]
Intra-tester (intra-session) reliability	ICC = 0.90 (0.83–0.94) Sample size = 30	Indeterminate	Did not pool the results or grade the evidence due to there being one study available [18]
Intra-tester (inter-session) reliability	ICC = 0.95 (0.91–0.97) Consistent results Sample size = 236	Sufficient	Low (as all studies are doubtful [11,56,60])
Prone Double Straight-Leg Test			
Inter-tester reliability	ICC = 0.83 (0.67–0.93) Sample size = 30	Indeterminate	Did not pool the results or grade the evidence due to there being one study available [18]
Intra-tester (intra-session) reliability	ICC = 0.86 (0.77–0.92) Sample size = 30	Indeterminate	Did not pool the results or grade the evidence due to there being one study available [18]

Commented [AT13]: Please define as appropriate.

Commented [AT14]: Please check the intended meaning is retained.

Commented [AT15]: Please check the intended meaning is retained.

Commented [AT16]: Please check the intended meaning is retained.

Commented [AT17]: Please check the intended meaning is retained.

Dynamic Extensor Endurance Test			
Intra-tester (inter-session) reliability	ICC = 0.99	Sufficient	Low (as there is one inadequate study [65])
	(0.88–1.00)		
	Consistent results		
Sample size: 82			

^a Pooled results obtained from mean reliability analysis and adjusted according to the publication bias analysis. ^b Overall rating was graded as sufficient (intraclass correlation coefficient (ICC) > 0.70), insufficient (ICC < 0.70), or indeterminate (either ICC reported by just one study or by none). ^c Quality of evidence (high, moderate, low, very low) based on the GRADE approach that uses three factors: (1) risk of bias assessed with the consensus-based standards for the selection of health measurement instruments (COSMIN) risk of bias check-list; (2) inconsistency, solved by pooling the results; and (3) imprecision, the total sample included in the studies. When the total sample size of the pooled studies is below 100, downgrade with one level and with two levels when the total sample size is below 50.

Commented [AT18]: Please define as appropriate.

3.4. Effect Sizes

3.4.1. Primary Outcomes

Table 2 presents the average reliability of the trunk extensor measures obtained from the different field-based tests and separately by type of reliability. The random effect models for the Biering-Sorensen test showed an average intra-tester and inter-session reliability of $ICC+ = 0.88$ (95% CI = 0.80 to 0.92, $I^2 = 88.9%$) obtained from 27 cohorts, an average intra-tester and intra-session reliability of $ICC+ = 0.88$ (95% CI = 0.83 to 0.92, $I^2 = 51.2%$) from 12 cohorts, and an average inter-tester reliability of $ICC+ = 0.94$ (95% CI = 0.84 to 0.98, $I^2 = 93.1%$) from eight cohorts. The prone isometric chest raise test reported an average intra-tester and inter-session reliability of $ICC+ = 0.95$ (95% CI = 0.91 to 0.97, $I^2 = 70.5%$) from five cohorts. Finally, the average intra-tester and inter-session reliability of the dynamic extensor endurance test was $ICC+ = 0.99$ (95% CI = 0.88 to 1.0, $I^2 = 74.5%$). Figure 2 displays a summary of the intra-tester and inter-session reliability obtained in the studies that applied the Biering-Sorensen test.

Table 2. Average reliability, 95% confidence intervals, and heterogeneity statistics for each field-based test and reliability type.

	<i>k</i>	<i>ICC+</i>	<i>ICC_t</i>	<i>ICC_u</i>	<i>Q</i>	<i>df</i>	<i>p</i>	<i>I²</i>
Biering-Sorensen Test:								
▪ Inter-tester reliability	8	0.94	0.84	0.98	102.37	7	<0.001	93.1%
▪ Intra-tester and intra-session reliability	12	0.88	0.83	0.92	22.69	11	0.02	51.2%
▪ Intra-tester and inter-session reliability	27	0.88	0.80	0.92	234.40	26	<0.001	88.9%
Prone Isometric Chest Raise test:								
▪ Inter-tester reliability	1	0.90	0.80	0.95	--	--	--	--
▪ Intra-tester and intra-session reliability	2	0.90	0.83	0.94	0.034	1	0.853	0%
▪ Intra-tester and inter-session reliability	5	0.95	0.91	0.97	13.57	4	0.009	70.5%
Prone Double Straight-Leg Test:								
▪ Inter-tester reliability	1	0.83	0.67	0.92	--	--	--	--
▪ Intra-tester and intra-session reliability	2	0.86	0.77	0.92	0.08	9	0.777	0%
Dynamic Extensor Endurance Test:								
▪ Intra-tester and inter-session reliability	5	0.99	0.88	1.00	15.73	4	0.003	74.5%

k: number of cohorts, *ICC*: mean intraclass correlation coefficient, *ICCL* and *ICCU*: 95% lower and upper CI for *ICC*, *Q*: heterogeneity statistic, *DF*: degrees of freedom for *Q* statistic, *p*: probability level associated to *Q* statistic, *I*²: heterogeneity index.

In the different meta-analyses carried out, the effect sizes exhibited a moderate to large heterogeneity (based on the *Q* statistics and the *I*² indices), supporting the decision of applying the random-effects model.

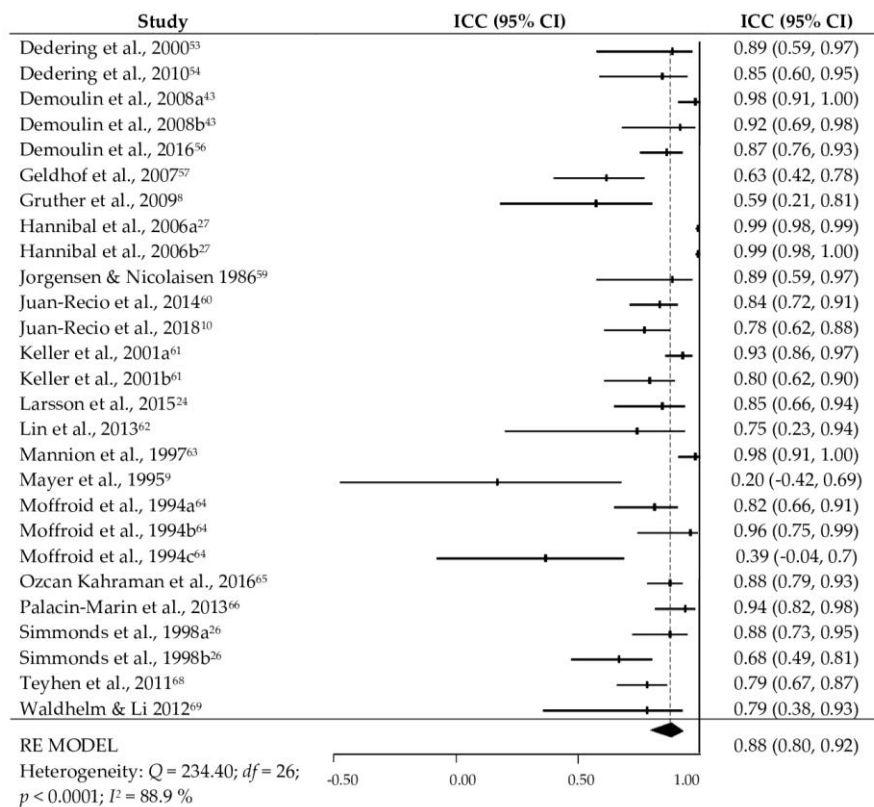


Figure 2. Forest plot of the intra-tester and inter-session reliability obtained in the studies that applied the Biering-Sorensen test.

Commented [M19]: Please changed the citation of references to the correct format. e.g., [1,2]...

3.4.2. Analysis of the Moderator Variables

As only the intra-tester and inter-session reliability for the Biering-Sorensen test presented more than 20 reliability estimates, the analyses of moderator variables were carried out exclusively for this field-based test and type of reliability, in order to examine the influence of qualitative and continuous moderator variables on the reliability coefficients.

Table 3 shows the results of the meta-regression analyses for each continuous moderator variable. None of the continuous moderators exhibited a statistically significant relationship with the intra-tester and the inter-session reliability of the Biering-Sorensen test.

Table 3. Results of the mixed-effects meta-regressions for the continuous moderator variables on the intraclass correlation coefficient (ICC) estimates obtained from intra-tester and inter-session reliability of the Biering-Sorensen test.

Moderator Variable	k	b _j	Q _R	p	Q _E	R ²
Publication year	27	0.003	0.03	0.855	232.21 *	0
Final sample size	27	-0.003	0.11	0.741	230.26 *	0
Sex (% female)	24	0.002	0.37	0.542	200.78 *	0
Mean age (years)	26	-0.017	1.62	0.203	210.81 *	0.01
SD age	19	-0.009	0.31	0.578	177.94 *	0
% attrition	27	-0.008	1.39	0.238	227.47 *	0
Number of measurements	27	-0.034	0.01	0.933	232.93 *	0
Time interval between measurement	26	-0.006	0.18	0.668	229.87 *	0
Mean test score from total sample	26	0.002	0.98	0.321	210.60 *	0
SD test score from total sample	21	0.009	0.94	0.331	201.42 *	0.02
Mean test score from reliability sample	24	0.001	0.42	0.514	207.27 *	0
SD test score from reliability sample	20	0.012	1.54	0.214	180.06 *	0.10

k: number of studies, b_j: regression coefficient for the predictor variable, Q_R: statistic for testing the statistical significance of the predictor variable, p: probability level for the Q_R statistic, Q_E: statistic for testing the model misspecification, R²: proportion of variance explained by the predictor variable, * p < 0.001.

Regarding the qualitative moderator variables, Table 4 shows the ANOVA results. Concerning the description of the field-based tests, significant statistical differences were obtained in position control instruments (p < 0.0001) with an explained variance of 69%. Specifically, the studies that used a plumb-line to control position exhibited, on average, greater reliability (ICC+ = 0.99), followed by the stadiometer (ICC+ = 0.89), the light sensor (ICC+ = 0.87), the inclinometer (ICC+ = 0.83), and the visual control (ICC+ = 0.81). Similarly, statistically significant differences were observed (p = 0.045) between the use of a roman chair (ICC+ = 0.94) or a test bench (ICC+ = 0.85), with an explained variance of 32%.

Table 4. Results of the mixed-effects ANOVAs for the qualitative moderator variables on the intraclass correlation coefficient (ICC) estimates obtained from intra-tester and inter-session reliability of the Biering-Sorensen test.

Moderator Variables	k	ICC+	95% CI		ANOVA Results
			ICCL	ICCU	
Reliability analysis was done with the same sample:					Q _B (1) = 0.92, p = 0.338; R ² = 0.01
Yes	24	0.88	0.81	0.92	Q _W (25) = 217.83, p < 0.0001
No	3	0.77	0.26	0.94	
▪ Sex:	9				Q _B (2) = 5.98, p = 0.050; R ² = 0.27
Males	2	0.88	0.77	0.94	Q _W (23) = 156.41, p < 0.0001
Females	15	0.97	0.89	0.99	
Males and females		0.83	0.72	0.89	Q _B (1) = 3.73, p = 0.053; R ² = 0.18
▪ Sample type:					Q _W (25) = 185.35, p < 0.0001
Children and adolescents	4	0.95	0.86	0.98	
Adults	23	0.85	0.77	0.90	Q _B (1) = 0.24, p = 0.625; R ² = 0
▪ Target population:					Q _W (25) = 231.60, p < 0.0001
Asymptomatic	18	0.88	0.80	0.93	
Clinical	9	0.85	0.69	0.93	Q _B (1) = 0.78, p = 0.332; R ² = 0.03
▪ Physical activity level:					Q _W (25) = 214.29, p < 0.0001
Sedentary	12	0.84	0.71	0.92	
Recreationally active	15	0.89	0.81	0.94	

<ul style="list-style-type: none"> ▪ Validated modification: 					$Q_B(1) = 0.24, p = 0.626; R^2 = 0$
Yes	17	0.86	0.76	0.92	
No	10	0.89	0.78	0.95	$Q_W(25) = 222.31, p < 0.0001$
<ul style="list-style-type: none"> ▪ Tool: 					$Q_B(1) = 4.03, p = 0.045; R^2 = 0.32$
Test bench	22	0.85	0.77	0.90	
Roman chair	5	0.94	0.86	0.98	$Q_W(25) = 158.11, p < 0.0001$
<ul style="list-style-type: none"> ▪ Hands position: 					$Q_B(2) = 0.37, p = 0.828; R^2 = 0$
Crossed on the chest	22	0.88	0.80	0.93	
Along the body	3	0.87	0.55	0.97	$Q_W(24) = 217.66, p < 0.0001$
At the level of the ears	2	0.80	0.23	0.96	
<ul style="list-style-type: none"> ▪ Part of the body on the edge: 					$Q_B(3) = 1.26, p = 0.7384; R^2 = 0$
Not reported	12	0.90	0.80	0.95	
ASIS	6	0.88	0.68	0.95	
Upper border of the iliac crest	6	0.85	0.62	0.95	$Q_W(23) = 212.02, p < 0.0001$
Pubis	3	0.77	0.28	0.94	
<ul style="list-style-type: none"> ▪ Test duration: 					$Q_B(1) = 1.24, p = 0.264; R^2 = 0.01$
Until exhaustion	25	0.88	0.82	0.93	
Until 240 s	2	0.72	0.08	0.94	$Q_W(25) = 212.46, p < 0.0001$
<ul style="list-style-type: none"> ▪ Position control instruments: 					$Q_B(4) = 34.49, p < 0.0001; R^2 = 0.69$
Visual	14	0.81	0.73	0.87	
Inclinometer	5	0.83	0.67	0.91	
Stadiometer	4	0.89	0.78	0.95	$Q_W(22) = 72.29, p < 0.0001$
Light sensor	2	0.87	0.59	0.96	
Plumb-line	2	0.99	0.98	0.99	
<ul style="list-style-type: none"> ▪ Familiarization session: 					$Q_B(1) = 12.09, p = 0.0005; R^2 = 0.34$
Yes	6	0.96	0.92	0.98	
No	20	0.82	0.73	0.88	$Q_W(25) = 151.63, p < 0.0001$
<ul style="list-style-type: none"> ▪ Test conditions: 					$Q_B(1) = 0.27, p = 0.605; R^2 = 0$
Similar conditions	18	0.88	0.80	0.93	
Unclear conditions	9	0.85	0.68	0.93	$Q_W(25) = 222.08, p < 0.0001$
<ul style="list-style-type: none"> ▪ The profession of tester: 					$Q_B(2) = 1.88, p = 0.389; R^2 = 0.02$
Sports sciences	6	0.92	0.80	0.97	
Physical therapy	17	0.86	0.76	0.92	$Q_W(23) = 205.35, p < 0.0001$
Medicine	3	0.77	0.30	0.94	
<ul style="list-style-type: none"> ▪ Continent: 					$Q_B(1) = 0.0004, p = 0.984; R^2 = 0$
Europe	16	0.88	0.78	0.93	
America	10	0.88	0.74	0.94	$Q_W(24) = 226.96, p < 0.0001$
<ul style="list-style-type: none"> ▪ Study objective: 					$Q_B(1) = 0.02, p = 0.877; R^2 = 0$
Psychometric	20	0.88	0.79	0.93	
Not psychometric	7	0.86	0.67	0.95	$Q_W(25) = 231.27, p < 0.0001$

<ul style="list-style-type: none"> ▪ Conflict of interest: <table border="0" style="margin-left: 20px;"> <tr> <td>No conflict</td> <td>8</td> <td>0.87</td> <td>0.71</td> <td>0.94</td> <td rowspan="2" style="vertical-align: middle;"> $Q_B(1) = 0.02, p = 0.896; R^2 = 0$ $Q_W(25) = 232.67, p < 0.0001$ </td> </tr> <tr> <td>Not reported</td> <td>19</td> <td>0.88</td> <td>0.79</td> <td>0.93</td> </tr> </table> ▪ COSMIN Risk of Bias <table border="0" style="margin-left: 20px;"> <tr> <td>check-list:</td> <td></td> <td></td> <td></td> <td></td> <td rowspan="5" style="vertical-align: middle;"> $Q_B(3) = 0.81, p = 0.848; R^2 = 0$ $Q_W(23) = 223.47, p < 0.0001$ </td> </tr> <tr> <td>Very good</td> <td>6</td> <td>0.85</td> <td>0.62</td> <td>0.94</td> </tr> <tr> <td>Adequate</td> <td>2</td> <td>0.80</td> <td>0.22</td> <td>0.96</td> </tr> <tr> <td>Doubtful</td> <td>17</td> <td>0.85</td> <td>0.79</td> <td>0.93</td> </tr> <tr> <td>Inadequate</td> <td>2</td> <td>0.92</td> <td>0.56</td> <td>0.99</td> </tr> </table> 	No conflict	8	0.87	0.71	0.94	$Q_B(1) = 0.02, p = 0.896; R^2 = 0$ $Q_W(25) = 232.67, p < 0.0001$	Not reported	19	0.88	0.79	0.93	check-list:					$Q_B(3) = 0.81, p = 0.848; R^2 = 0$ $Q_W(23) = 223.47, p < 0.0001$	Very good	6	0.85	0.62	0.94	Adequate	2	0.80	0.22	0.96	Doubtful	17	0.85	0.79	0.93	Inadequate	2	0.92	0.56	0.99
No conflict	8	0.87	0.71	0.94	$Q_B(1) = 0.02, p = 0.896; R^2 = 0$ $Q_W(25) = 232.67, p < 0.0001$																																
Not reported	19	0.88	0.79	0.93																																	
check-list:					$Q_B(3) = 0.81, p = 0.848; R^2 = 0$ $Q_W(23) = 223.47, p < 0.0001$																																
Very good	6	0.85	0.62	0.94																																	
Adequate	2	0.80	0.22	0.96																																	
Doubtful	17	0.85	0.79	0.93																																	
Inadequate	2	0.92	0.56	0.99																																	

k: number of cohorts, ASIS: anterosuperior iliac spine, ANOVA: analysis of variance, ICC: mean reliability coefficient, ICC_L and ICC_U: lower and upper 95% confidence limits for ICC, Q_B: statistic for testing the statistical significance of the predictor variable, Q_W: statistic for testing the model misspecification, R²: proportion of variance explained by the predictor variable.

Concerning the characteristics of the study design, the presence of a familiarization session before data collection showed a statistically significant relationship with the reliability estimates ($p = 0.0005$). Specifically, carrying out a familiarization session showed larger average reliability ($ICC+ = 0.96$) than when it was not performed ($ICC+ = 0.82$), with an explained variance of 34%.

Finally, the sex of the sample approached statistical significance on the average reliability ($p = 0.05, R^2 = 0.27$), with the female samples presenting a better reliability ($ICC+ = 0.97$) than the males ($ICC+ = 0.88$) or males and females together ($ICC+ = 0.83$). Similarly, the sample type showed a tendency towards statistical significance ($p = 0.053, R^2 = 0.18$). The studies that used children or adolescents as the sample showed, on average, higher reliability ($ICC+ = 0.95$) than those that tested adults ($ICC+ = 0.85$).

3.5. An Explanatory Model

As a further step, a multiple meta-regression was applied to identify the subset of moderator variables that can explain most of the reliability coefficient variability. The predictors included in the model were selected on the basis of the ANOVAs and meta-regression results previously conducted. Thus, out of the five moderator variables that presented a statistically significant relationship with the reliability coefficients (sex, sample type, tool, position control instruments, and familiarization session), a multiple regression model with two predictors was the model that best explained the variability of the reliability coefficients—the type of sample (0, adults vs. 1, children/adolescents) and the familiarization session (0, No vs. 1, Yes). Table 5 presents the results. These two moderator variables reached statistical significance ($Q_R(2) = 20.57, p < 0.001$) with 51% of variance accounted for. As shown in Table 4, both moderator variables exhibited a statistically significant relationship with the reliability coefficients, once the influence of the other moderator was controlled.

Table 5. Results of the multiple meta-regression model assuming a mixed-effects model.

Source	b_i	SE	Z	p
Intercept	1.09	0.11	9.89	<0.001
Sample type	0.65	0.27	2.39	0.017
Familiarization session	0.68	0.24	2.88	0.004
Full model:			$Q_R(2) = 20.57, R^2 = 0.51$	
			$Q_E(24) = 113.67, p < 0.001$	

b_i: unstandardized regression coefficient, SE: standard error of *b_i*, Z: statistic for testing the statistical significance of each moderator variable, Q_R: statistic for testing the global significance of the model, Q_E: statistic for testing the model misspecification, R²: proportion of variance accounted for by the full model.

3.6. Reliability Induction

Out of the 103 empirical studies that reported some reliability coefficients, 75 studies induced reliability (37 by omission and 38 by the report from other studies), which implies a 72.8% of RI. In particular, of these 38 studies that induced reliability by the report from other studies (41 cohorts), 32 induced reliability of the Biering-Sorensen test, 4 the prone isometric chest raise test, 2 the supine bridge test, 1 the prone double straight-leg raise test, and 2 the reliability of the dynamic extensor endurance test (see the references in Document S5 in the Supplementary Materials section).

4. Discussion

The main purpose of the current RG meta-analysis was to estimate both the inter- and intra-tester (intra- and inter-session) reliability of the trunk extension endurance measures obtained through four field-based tests (Biering-Sorensen, prone isometric chest raise, prone double straight-leg raise, and dynamic extensor endurance), as well as to identify those characteristics (qualitative and quantitative moderators) of the studies selected that might have a meaningful influence in the degree of heterogeneity showed by the pooled reliability coefficients (ICC scores). A secondary purpose was to determine the RI practice in studies that used at least one of the four trunk muscle endurance field-based tests investigated in this RG meta-analysis.

The systematic literature review carried out showed that very few studies have analyzed the inter-tester reliability of the trunk extension endurance measures obtained through field-based tests [18,23,24,26,66]. Specifically, most of these studies focused on the analysis of the inter-tester reliability of the trunk extension endurance measure obtained through the Biering-Sorensen test [18,23,24,26,66] (eight cohorts); the number of studies that explored the inter-tester reliability of the endurance measures recorded from the prone isometric chest raise [18] (one cohort) and the prone double straight-leg [18] (one cohort) tests were very limited, and no studies used the dynamic extensor endurance test. In particular, the results of this RG meta-analysis showed appropriate inter-tester reliability values ($ICC+ > 0.80$) for the trunk extension endurance measures obtained through the Biering-Sorensen ($ICC+ = 0.94$), the prone chest raise ($ICC+ = 0.90$), and the prone double straight-leg ($ICC+ = 0.83$) tests. However, and due to the scarce evidence available (one cohort with a small sample size ($n = 30$) was found), the inter-tester reliability estimates of the measures obtained through the prone isometric chest raise and prone double straight-leg tests should be considered with a degree of caution.

On the other hand, five studies [18,25,26,43,67] analyzed the intra-tester reliability of the trunk extension endurance measures, using short periods (intra-session, within a day). Most of these studies focused exclusively on the endurance measure obtained through the Biering-Sorensen test, which reported appropriate reliability results ($ICC+ = 0.88$). For the rest of the trunk extensor endurance field-based tests, only one study was found that assessed the intra-tester reliability in its intra-session modality for the prone isometric chest raise and the prone double straight-leg tests [18], reporting acceptable reliability results ($ICC+ = 0.90$ and 0.86 , respectively). However, these intra-tester reliability results for the prone isometric chest raise and the prone double straight-leg tests were not supported by strong evidence (only 1 cohort of 30 participants [18]) and hence should be considered with a degree of caution. There appears to be no study focusing on the intra-session reliability of the measure obtained from the dynamic extensor endurance test.

Contrarily, there was a higher number of studies that analyzed the intra-tester and inter-session reliability of the trunk extension endurance measures obtained through the field-based tests selected (except the endurance measures obtained from the prone double straight-leg test, as no study has analyzed its inter-session reliability). Different authors [70,71] have suggested that the scarce number of intra-session reliability coefficients reported in the literature may be explained by the fact that field-based tests are often used in longitudinal randomized controlled trials that analyze the chronic effects of training interventions rather than in studies that were conducted using short periods to analyze the acute effects of specific interventions. Therefore, the higher number of studies that have analyzed the inter-session reliability in contrast with those that have analyzed the intra-session reliability of the endurance measures seems to be in line with the research designs and interests currently present in the scientific literature. All the trunk extension measures analyzed reported

appropriate inter-session reliability scores, with ICC values ranging from $ICC+ = 0.88$ (Biering-Sorensen test) to 0.99 (dynamic extensor endurance test).

On the whole, only the trunk extension endurance measure obtained through the Biering-Sorensen test may present sufficient scientific evidence in terms of inter- ($ICC+ = 0.94$, based on 8 cohorts) and intra-tester (intra-session ($ICC+ = 0.88$, based on 12 cohorts) and inter-session ($ICC+ = 0.88$, based on 27 cohorts)) reliability to justify its use for practical and clinical purposes.

The analysis of the quantitative and qualitative factors or moderators that may have had an influence in the degree of heterogeneity found for the pooled reliability coefficients of the trunk extension endurance measures could only be carried out in the measurements obtained through the Biering-Sorensen test and for the inter-session subtype of intra-tester reliability because it was the only one with at least 20 reported reliability estimates. Thus, the results derived from the meta-regressions demonstrated that none of the factors/objects of the study had a statistically significant impact ($p > 0.05$) on the heterogeneity exhibited by the pooled inter-session reliability coefficient of the endurance measure obtained through the Biering-Sorensen test. It should be highlighted that even though a non-statistically significant relationship was found between the moderator "SD test score from reliability sample" and the inter-session reliability coefficient of the Biering-Sorensen test, it explained 10% of the variance. This finding agrees with the psychometric theory that states that the larger the SD of test scores, the greater the reliability that might be obtained [72].

Regarding the qualitative moderator analysis (ANOVAs), the instrument or tool used to control the participant's position during the test (no instrument (visually controlled) vs. inclinometer vs. stadiometer vs. light sensor vs. plumb-line), the equipment used to perform the test (test bench vs. roman chair) and the presence (or not) of a familiarization session were the moderators that showed statistically significant effects ($p < 0.05$) on the inter-session reliability estimate of the endurance measure obtained through the Biering-Sorensen test.

During the execution of the Biering-Sorensen test, small movements sometimes occur naturally while participants are trying to maintain the required isometric horizontal position. These little movements may make the tester falsely assume a perception of fatigue in the participant who is being tested [23]. To minimize this source of error, it has been recommended that the tester should control the participant's position during the test through the use of a static visual reference. The results of this RG meta-analysis showed that the use of a plumb-line to control the participant's position (probably because, and unlike the other tools that only offer visual feedback of the participant's position to the tester, it also provides kinesthetic feedback in the low back region to the participant who is being tested about the position that has to be maintained during the test) was the instrument that reported the highest inter-session reliability ($ICC+ = 0.99$), followed by the stadiometer ($ICC+ = 0.89$), light sensor ($ICC+ = 0.87$), inclinometer ($ICC+ = 0.83$), and visual control ($ICC+ = 0.81$).

The tool or equipment used to perform the Biering-Sorensen test also showed significant effects on the inter-session reliability estimate. Those studies that used the roman chair to perform the Biering-Sorensen test reported better inter-session reliability scores ($ICC+ = 0.94$) than those using the test bench ($ICC+ = 0.85$). However, this difference might be partially attributed to the imbalance distribution existing regarding the number of cohorts selected to carry out the inter-session reliability analysis of both sub-categories of this moderator. Thus, whereas five cohorts were used to calculate the inter-session reliability estimate of the trunk extensor endurance measure obtained through the Biering-Sorensen test using a roman chair, 22 cohorts were used to conduct the same analysis using a test bench.

One of the main sources of error in reliability studies is the presence of learning effects [14]. This effect was identified, as in studies in which participants were asked to carry out a familiarization session for the Biering-Sorensen test before the data collection reported statistically significantly better inter-session reliability coefficients ($ICC+ = 0.96$) than those that did not employ a familiarization session ($ICC+ = 0.82$). The use of a familiarization session before the assessment of the trunk extensor endurance may help the participant to learn, for example, the appropriate testing position needed to be adopted for the test and to better tolerate the fatigue feelings generated during the Biering-Sorensen test [10,56]. However, quite often it is impossible to carry out a familiarization

Commented [AT20]: Please check the intended meaning is retained.

session in practical (clinic, sports, fitness, and physical education) contexts. Although this circumstance affects the inter-session reliability estimate of the Biering-Sorensen test negatively, the magnitude of this effect is not large enough to consider this endurance measure inappropriate (according to the 0.8 cut-off score) for practical and research purposes.

Other qualitative moderators such as sex (males vs. females vs. males and females) and population of sample (children and adolescents vs. adults) also showed a tendency towards statistical significance. However, and similar to what was found for the moderator tool or equipment, these tendencies may be caused by the large imbalanced distribution shown in the number of cohorts included in each of the sub-categories of these moderators (e.g., males = 9 cohorts vs. females = 2 cohorts vs. males and females = 15 cohorts; adults = 23 cohorts vs. children and adolescents = 4 cohorts). Despite this, the sub-categories of these two moderators showed reliability estimates higher than 0.8, and hence the Biering-Sorensen test may be a suitable field-based test to be conducted in both sexes and youth and adult populations.

Finally, out of 103 studies that reported reliability coefficients, 75 studies induced reliability (72.8%). The number of studies that induced reliability according to our study is in concordance with the high rate of RI that usually characterizes research in general and which is around 70% of the studies conducted in the social and health sciences [73]. These findings reinforce the need to educate researchers that reliability is not an inherent characteristic of the instrument and that it should be analyzed whenever a test is administered.

5. Limitations

This RG meta-analysis presents some limitations that should be highlighted. First, the unadvised practice of inducing reliability reduced the number of studies included in this RG meta-analysis. Second, only studies in English or Spanish were included, which may have limited the number of potentially eligible studies. Third, the lack of important data reported by authors reduced the possibility of analyzing their influence as potential moderating variables on the reliability coefficients. Such was the case of the age means and age standard deviation of the samples, as well as the sex distribution, the physical activity level, or the tester's profession and level of experience conducting the test. To finish, the scarce number of reliability estimates reported for the field tests other than the Biering-Sorensen test did not allow to analyze the moderator variables that may explain in part the variability found among the reliability coefficients for those tests.

6. Conclusions

The main findings of the current RG meta-analysis report that only the trunk extension endurance measure obtained through the Biering-Sorensen test may present sufficient scientific evidence in terms of inter- ($ICC+ = 0.94$, based on 8 cohorts) and intra-tester (intra-session ($ICC+ = 0.88$, based on 12 cohorts) and inter-session ($ICC+ = 0.88$, based on 27 cohorts)) reliability to justify its use (mainly in male adults) for practical and clinical purposes. Furthermore, this trunk extension endurance measure may be even more reliable when a familiarization session with the Biering-Sorensen testing procedure is previously carried out and a plumb-line is used to control the participant's position during the test.

None of the rest of the trunk extension endurance measures obtained through the other static (prone isometric chest raise test and prone double straight-leg raise test) and dynamic field-based (dynamic extensor endurance test) tests showed sufficient scientific evidence (in terms of reliability) to promote their use for practical and clinical purposes. Therefore, more reliability studies conducted in different cohorts are needed before the use of these field-based tests can be recommended to assess trunk extensor endurance in both practical and research settings.

Finally, the high RI rate found in this study (72.8%) suggests that researchers should be more aware of the fact that RI is an erroneous practice that should be eradicated because it can cause errors in the estimation of the measures used [74]. Thus, in future research, authors should report the reliability coefficient of the scores for the data being analyzed, even when the focus of their research is not psychometric.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Table S1: PRISMA Checklist. Table S2: Search strategy. Table S3: Description of the COSMIN methodology. Table S4: Individual characteristics of each study selected. Table S5: COSMIN risk of bias checklist (Box 6 reliability) and criteria for good measurement properties. Box S1: Coding protocol. Document S1: References of articles that practice RI.

Author Contributions: Conceptualization, M.T.M.-R. and F.A.; methodology, F.A. and J.S.-M.; software, J.S.-M.; validation, P.S.d.B. and F.S.-M.; formal analysis, J.S.-M.; investigation, M.T.M.-R., F.S.-M., and P.S.d.B.; resources, M.T.M.-R. and J.S.-M.; data curation, M.T.M.-R. and P.S.d.B.; writing—original draft preparation, M.T.M.-R.; writing—review and editing, F.A., M.D.S.C., and F.J.V.-G.; visualization, F.A., M.D.S.C., and F.J.V.-G.; supervision, J.S.-M., P.S.d.B., and F.S.-M.; project administration, P.S.d.B. and F.S.-M.; funding acquisition, M.T.M.-R. and P.S.d.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a predoctoral grant awarded by the Ministerio de Educación, Cultura y Deporte (grant number: FPU15/05200). This study is part of the project entitled "Study of risk of injury in young athletes through artificial intelligence networks", funded by Ministerio de Ciencia e Innovación (DEP2017-88775-P). This study was carried out during the research stay at the School of Sport and Exercise of the University of Gloucestershire from 01/05/2018 to 01/08/2018 funded by the Ministerio de Educación, Cultura y Deporte.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

- Abdelraouf, O.R.; Abdel-aziem, A.A. The relationship between core endurance and back dysfunction in collegiate male athletes with and without nonspecific low back pain. *Int. J. Sports Phys. Ther.* **2016**, *11*, 337–344.
- Biering-Sorensen, F. Physical Measurements as Risk Indicators of Low-Back trouble over one-year period. *Spine (Phila Pa 1976)* **1984**, *9*, 106–119.
- Luoto, S.; Heliövaara, M.; Hurri, H.; Alaranta, H. Static back endurance low-back pain and the risk of low-back pain. *Clin. Biomech.* **1995**, *10*, 323–324.
- McGill, S.M.; Childs, A.; Liebenson, C. Endurance times for low back stabilization exercises: Clinical targets for testing and training from a normal database. *Arch. Phys. Med. Rehabil.* **1999**, *80*, 941–944.
- Hibbs, A.E.; Thompson, K.G.; French, D.; Wrigley, A.; Spears, I. Optimizing performance by improving core stability and strength. *Sport Med.* **2008**, *38*, 995–1008.
- McGill, S.M. *Ultimate back Fitness and Performance*, 4th ed.; Backfitpro Incorporated: Waterloo, ON, Canada, 2009.
- García-Vaquero, M.P.; Barbado, D.; Juan-Recio, C.; López-Valenciano, A.; Vera-García, F.J. Isokinetic trunk flexion-extension protocol to assess trunk muscle strength and endurance: Reliability, learning effect, and sex differences. *J. Sport Health Sci.* **2016**, doi:10.1016/j.jshs.2016.08.011.
- Gruther, W.; Wick, F.; Paul, B.; Leitner, C.; Posch, M.; Matzner, M.; Crevena, R.; Ebenbichler, G. Diagnostic accuracy and reliability of muscle strength and endurance measurements in patients with chronic low back pain. *J. Rehabil. Med.* **2009**, *41*, 613–619, doi:10.2340/16501977-0391.
- Mayer, T.; Gatchel, R.; Betancur, J.; Bovasso, E. Trunk Muscle Endurance Measurement. Isometric Contrasted to Isokinetic Testing in Normal Subjects. *Spine (Phila Pa 1976)* **1995**, *20*, 920–927.
- Juan-Recio, C.; López-Plaza, D.; Barbado, D.; García-Vaquero, M.P.; Vera-García, F.J. Reliability assessment and correlation analysis of 3 protocols to measure trunk muscle strength and endurance muscle strength and endurance. *J. Sports Sci.* **2018**, *36*, 357–364, doi:10.1080/02640414.2017.1307439.
- Ito, T.; Shirado, O.; Suzuki, H.; Takahashi, M.; Kaneda, K.; Strax, T.E. Lumbar trunk muscle endurance testing: An inexpensive alternative to a machine for evaluation. *Arch. Phys. Med. Rehabil.* **1996**, *77*, 75–79.
- McIntosh, G.; Wilson, L.; Affleck, M.; Hall, H. Trunk and lower extremity muscle endurance: Normative data. *J. Rehabil. Outcomes Meas.* **1998**, *2*, 20–39.
- Lanning, C.L.; Uhl, T.L.; Ingram, C.L.; Mattacola, C.G.; English, T.; Newsom, S. Baseline values of trunk endurance and hip strength in collegiate athletes. *J. Athl. Train.* **2006**, *41*, 427–434.
- Hopkins, W.G. Measures of Reliability in Sports Medicine and Science. *Sports Med.* **2000**, *30*, 375–381, doi:10.2165/00007256-200030010-00001.
- Riebe, D.; Ehrman, J.K.; Liguori, G.; Magal, M. *ACSM's Guidelines for Exercise Testing and Prescription*, 10th ed.; Wolters Kluwer: Philadelphia, PA, USA, EEUU, 2018.

Commented [M21]: 1. Ref 15, 46: please confirm if it should be deleted.

2. Ref 52: please add location or provide the website URL and accessed date to us.

3. Ref 68: please add volume or doi.

4. Ref 73: please add publisher and city, or provide the website URL and accessed date to us.

16. Maier, T.; Gross, M.; Trösch, S.; Steiner, T.; Müller, B.; Bourban, P.; Schärer, C.; Hübner, K.; Wehrli, J.; Tschopp, M. *Manual Leistungsdiagnostik*; BASPO: Magglingen, Switzerland, 2016.
17. Plowman, S.A.; Meredith, M.D. *FITNESSGRAM®/ACTIVITYGRAM®. Reference Guide*, 4th ed.; The Cooper Institute: Dallas, TX, USA, 2013.
18. Arab, M.A.; Salavati, M.; Ebrahimi, I.; Mousavi, M.E. Sensitivity, specificity and predictive value of the clinical trunk muscle endurance tests in low back pain. *Clin. Rehabil.* **2007**, *21*, 640–647, doi:10.1177/0269215507076353.
19. Moreau, C.E.; Green, B.N.; Johnson, C.D.; Moreau, S.R. Isometric back extension endurance tests: A review of the literature. *J. Manip. Physiol. Ther.* **2001**, *24*, 110–122 doi:10.1067/mmt.2001.112563.
20. Atkinson, G.; Nevill, A. Statistical Methods for Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine. *Sports Med.* **1998**, *26*, 217–238, doi:10.2165/00007256-199826040-00002.
21. Kottner, J.; Audigé, L.; Brorson, S.; Donner, A.; Gajewski, B.J.; Hróbjartsson, A.; Roberts, C.; Shoukri, M.; Streiner, D.L. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J. Clin. Epidemiol.* **2011**, *64*, 96–106, doi:10.1016/j.jclinepi.2010.03.002.
22. Hayen, A.; Dennis, R.J.; Finch, C.F. Determining the intra- and inter-observer reliability of screening tools used in sports injury research. *J. Sci. Med. Sports* **2007**, *10*, 201–210, doi:10.1016/j.jsams.2006.09.002.
23. Latimer, J.; Maher, C.G.; Refshauge, K.; Colaco, I. The reliability and validity of the Biering-Sorensen test in asymptomatic subjects and subjects reporting current or previous nonspecific low back pain. *Spine (Phila Pa 1976)* **1999**, *24*, 2085–2089, doi:10.1097/00007632-199910150-00004.
24. Larsson, H.; Tegern, M.; Monnier, A.; Skoglund, J.; Helander, C.; Persson, E.; Malm, C.; Broman, L.; Aasa, U. Content Validity Index and Intra- and Inter-Rater Reliability of a New Muscle Strength/Endurance Test Battery for Swedish Soldiers. *PLoS ONE* **2015**, *10*, e0132185, doi:10.1371/journal.pone.0132185.
25. Moffroid, M.; Haugh, L.; Haig, A.; Henry, S.; Pope, M. Endurance training of trunk extensor muscles. *Phys. Ther.* **1993**, *73*, 10–17.
26. Simmonds, M.J.; Olson, S.L.; Jones, S.; Hussein, T.; Lee, E.; Novy, D.; Radwan, H. Psychometric Characteristics and Clinical Usefulness of Physical Performance Tests in Patients with Low Back Pain. *Spine (Phila Pa 1976)* **1998**, *23*, 2412–2421.
27. Hannibal, N.S., III; Plowman, S.A.; Looney, M.A.; Brandenburg, J. Reliability and Validity of Low Back Strength/Muscular Endurance Field Tests in Adolescents. *J. Phys. Act. Health* **2006**, *3*(Suppl. S2), 78–89, DOI:10.1123/jpah.3.s2.s78.
28. Balthazard, P.; de Goumoens, P.; Rivier, G.; Demeulenaere, P.; Ballabeni, P.; Deriaz, O. Manual therapy followed by specific active exercises versus a placebo followed by specific active exercises on the improvement of functional disability in patients with chronic non specific low back pain: A randomized controlled trial. *BMC Musculoskelet. Disord.* **2012**, *13*, 162, doi:10.1186/1471-2474-13-162.
29. Coorevits, P.; Danneels, L.; Cambier, D.; Ramon, H.; Vanderstraeten, G. Assessment of the validity of the Biering-Sorensen test for measuring back muscle fatigue based on EMG median frequency characteristics of back and hip muscles. *J. Electromyogr. Kinesiol.* **2008**, *18*, 997–1005, doi:10.1016/j.jelekin.2007.10.012.
30. Kankaanpää, M.; Laaksonen, D.; Taimela, S.; Kokko, S.M.; Airaksinen, O.; Hänninen, O. Age, sex, and body mass index as determinants of back and hip extensor fatigue in the isometric Sorensen back endurance test. *Arch. Phys. Med. Rehabil.* **1998**, *79*, 1069–1075, doi:10.1016/s0003-9993(98)90173-3.
31. Evans, K.; Refshauge, K.; Adams, R. Trunk muscle endurance tests: Reliability, and gender differences in athletes. *J. Sci. Med. Sports* **2007**, *10*, 447–455, doi:10.1016/j.jsams.2006.09.003.
32. Okada, T.; Huxel, K.C.; Nesser, T.W. Relationship between core stability, functional movement, and performance. *J. Strength Cond. Res.* **2011**, *25*, 252–261, doi:10.1519/JSC.0b013e3181b22b3e.
33. Vacha-Haase, T.; Kogan, L.R.; Thompson, B. Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educ. Psychol. Meas.* **2000**, *60*, 509–522, doi:10.1177/00131640021970682.
34. Henson, R.K.; Thompson, B. Characterizing measurement error in scores across studies: Some recommendations for conducting ‘reliability generalization’ studies. *Meas. Eval. Couns. Dev.* **2002**, *35*, 113–126, doi:10.1080/07481756.2002.12069054.
35. Rodriguez, M.C.; Maeda, Y. Meta-analysis of coefficient alpha. *Psychol. Methods* **2006**, *11*, 306–322, doi:10.1037/1082-989X.11.3.306.

36. Sánchez-Meca, J.; López-López, J.A.; López-Pina, J.A. Some recommended statistical analytic practices when reliability generalization studies are conducted. *Br. J. Math. Stat. Psychol.* **2013**, *66*, 402–425, doi:10.1111/j.2044-8317.2012.02057.x.
37. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann. Intern. Med.* **2009**, *151*, 264–269, doi:10.1371/journal.pmed.1000097.
38. Prinsen, C.A.C.; Mokkink, L.B.; Bouter, L.M.; Alonso, J.; Patrick, D.L.; de Vet, H.C.W.; Terwee, C.B. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual. Life Res.* **2018**, *27*, 1147–1157, doi:10.1007/s11136-018-1798-3.
39. Mokkink, L.B.; de Vet, H.C.W.; Prinsen, C.A.C.; Patrick, D.L.; Alonso, J.; Bouter, L.M.; Terwee, C.B. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual. Life Res.* **2018**, *27*, 1171–1179, doi:10.1007/s11136-017-1765-4.
40. Terwee, C.B.; Prinsen, C.A.C.; Chiarotto, A.; Westerman, M.; Patrick, D.; Alonso, J.; Bouter, L.M.; de Vet, H.C.W.; Mokkink, L.B. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Qual. Life Res.* **2018**, *27*, 1159–1170, doi:10.1007/s11136-018-1829-0.
41. Weir, J.P. Quantifying test-retest reliability using the intraclass correlation coefficient and the standard error of measurement. *J. Strength Cond. Res.* **2005**, *19*, 231–240, doi:10.1519/15184.1.
42. Salminen, J.J.; Maki, P.; Oksanen, A.; Pentti, J. Spinal mobility and trunk muscle strength in 15-year-old schoolchildren with and without low-back pain. *Spine (Phila Pa 1976)* **1992**, *17*, 405–411, doi:10.1097/00007632-199204000-00005.
43. Demoulin, C.; Sac, D.; Serre, L.; Maquet, D.; Crielaard, J.M.; Marc, V. Reproducibility and Suitability of Clinical Assessments of Trunk Flexor and Extensor Muscles. *J. Musculoskelet. Pain* **2008**, *16*, 298–308, doi:10.1080/10582450802479651.
44. Sandrey, M.A.; Mitzel, J.G. Improvement in Dynamic Balance and Core Endurance after a 6-Week Core-Stability-Training Program in High School Track and Field Athletes. *J. Sports Rehabil.* **2013**, *22*, 264–271, doi:10.1123/jsr.22.4.264.
45. Terwee, C.B.C.; Mokkink, L.; Knol, D.; Ostelo, R.; Bouter, L.; de Vet, H. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Qual. Life Res.* **2012**, *21*, 651–657, doi:10.1007/s11136-011-9960-1.
46. Orwin, R.; Vevea, J. Evaluating coding decisions. In *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed.; Cooper, H., Hedges, L.V., Valentine, J.C., Eds.; Russel Sage Foundation: New York, NY, USA, EEUU, 2009; pp. 177–203.
47. Borenstein, M.; Hedges, L.V.; Higgins, J.; Rothstein, H.R. *Introduction to Meta-Analysis*; John Wiley & Sons: Chichester, UK, 2009.
48. Sánchez-Meca, J.; Marín-Martínez, F. Confidence Intervals for the Overall Effect Size in Random-Effects Meta-Analysis. *Psychol. Methods* **2008**, *13*, 31–48, doi:10.1037/1082-989X.13.1.3.
49. Huedo-Medina, T.B.; Sánchez-Meca, J.; Marín-Martínez, F.; Botella, J. Assessing heterogeneity in meta-analysis: Q statistic or I² Index? *Psychol. Methods* **2006**, *11*, 193–206, doi:10.1037/1082-989X.11.2.193.
50. Lopez-Lopez, J.A.; Botella, J.; Sanchez-Meca, J.; Marín-Martínez, F. Alternatives for Mixed-Effects Meta-Regression Models in the Reliability Generalization Approach: A Simulation Study. *J. Educ. Behav. Stat.* **2013**, *38*, 443–469, doi:10.3102/1076998612466142.
51. López-López, J.A.; Marín-Martínez, F.; Sánchez-Meca, J.; Van den Noortgate, W.; Viechtbauer, W. Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *Br. J. Math. Stat. Psychol.* **2014**, *67*, 30–48, doi:10.1111/bmsp.1200.
52. Borenstein, M.; Hedges, L.V.; Higgins, J.; Rothstein, H.R. *Comprehensive Meta-Analysis 3.3*; Biostat, Inc: Publisher location, Country, 2014.
53. Dederig, A.; Roos Af Hjelmsäter, M.; Elfving, B.; Harms-Ringdahl, K.; Németh, G. Between-days reliability of subjective and objective assessments of back extensor muscle fatigue in subjects without lower-back pain. *J. Electromyogr. Kinesiol.* **2000**, *10*, 151–158, doi:10.1016/s1050-6411(00)00009-2.
54. Dederig, A.; Gnospelius, A.; Elfving, B. Reliability of Measurements of Endurance Time, Electromyographic Fatigue and Recovery, and Associations to Activity Limitations, in Patients with Lumbar Disc Herniation. *Physiother. Res. Int.* **2010**, *15*, 189–198, doi:10.1002/pri.457.

55. del Pozo-Cruz, B.; Mocholi, M.; del Pozo-Cruz, J.; Parraca, J.; Adsuar, J.; Gusi, N. Reliability and validity of lumbar and abdominal trunk muscle endurance tests in office workers with nonspecific subacute low back pain. *J. Back Musculoskelet. Rehabil.* **2014**, *27*, 399–408, doi:10.3233/BMR-140460.
56. Demoulin, C.; Boyer, M.; Duchateau, J.; Grosdent, S.; Jidovstev, B.; Crielaard, J.M.; Vanderthommen, M. Is the Sørensen test valid to assess muscle fatigue of the trunk extensor muscles? *J. Back Musculoskelet. Rehabil.* **2016**, *29*, 31–40, doi:10.3233/BMR-150592.
57. Geldhof, E.; Cardon, G.; De Bourdeaudhuij, I.; Danneels, L.; Coorevits, P.; Vanderstraeten, G.; de Clercq, D. Effects of back posture education on elementary schoolchildren's back function. *Eur. Spine J.* **2007**, *16*, 829–839, doi:10.1007/s00586-006-0199-4.
58. Javadian, Y.; Behtash, H.; Akbari, M.; Taghipour-darzi, M.; Zekavat, H. The effects of stabilizing exercises on pain and disability of patients with lumbar segmental instability. *J. Back Musculoskelet. Rehabil.* **2012**, *25*, 149–155, doi:10.3233/BMR-2012-0321.
59. Jorgensen, K.; Nicolaisen, T. Two methods for determining trunk extensor endurance. A comparative study. *Eur. J. Appl. Physiol. Occup. Physiol.* **1986**, *55*, 639–644, doi:10.1007/bf00423210.
60. Juan-Recio, C.; Barbado, D.; López-Valenciano, A.; Vera-García, F.J. Test de campo para valorar la resistencia de los músculos del tronco. *Apunt. Educ. Fis. Deport.* **2014**, *3*, 59–68, doi:10.5672/apunts.2014-0983.es.(2014/3).117.06.
61. Keller, A.; Hellesnes, J.; Brox, J.I.I. Reliability of the isokinetic trunk extensor test, Biering-Sorensen test, and Astrand bicycle test: Assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. *Spine (Phila Pa 1976)* **2001**, *26*, 771–777, doi:10.1097/00007632-200104010-00017.
62. Lin, K.-H.; Huang, Y.-M.; Tang, W.; Chang, Y.; Liu, Y.; Liu, C. Correlation of static and dynamic trunk muscle endurance and bat swing velocity in high school aged baseball players. *Isokinet. Exerc. Sci.* **2013**, *21*, 113–119, doi:10.3233/IES-130486.
63. Mannion, A.F.; Connolly, B.; Wood, K.; Dolan, P. The use of surface EMG power spectral analysis in the evaluation of back muscle function. *J. Rehabil. Res. Dev.* **1997**, *34*, 427–439.
64. Moffroid, M.; Reid, S.; Henry, S.; Haugh, L.; Ricamato, A. Some Endurance Measures in Persons With Chronic Low Back Pain. *J. Orthop. Sports Phys. Ther.* **1994**, *20*, 81–87, doi:10.2519/jospt.1994.20.2.81.
65. Ozcan Kahraman, B.; Sengul, Y.; Kahraman, T.; Kalemci, O. Developing a Reliable Core Stability Assessment Battery for Patients With Nonspecific Low Back Pain. *Spine (Phila Pa 1976)* **2016**, *41*, E844–E850, doi:10.1097/BRS.0000000000001403.
66. Palacin-Marin, F.; Esteban-Moreno, B.; Olea, N.; Herrera-Viedma, E.; Arroyo-Morales, M. Agreement between telerehabilitation and face-to-face clinical outcome assessments for low back pain in primary care. *Spine (Phila Pa 1976)* **2013**, *38*, 947–952, doi:10.1097/BRS.0b013e318281a36c.
67. de Souza, C.P.; Monteiro-Junior, R.S.; da Silva, E.B. Reliability of the endurance test for the erector spinae muscle. *Fisioter. Mov.* **2016**, *29*, 369–375, doi:10.1590/0103-5150.029.002.AO16.
68. Teyhen, D.S.; Shaffer, S.W.; Lorensen, C.L.; Wood, S.L.; Rogers, S.M.; Dugan, J.L.; Walker, M.J.; Childs, J.D. Reliability of lower quarter physical performance measures in healthy service members. *US Army Med. Dep. J.* **2011**, 37–49.
69. Waldhelm, A.; Li, L. Endurance tests are the most reliable core stability related measurements. *J. Sports Health Sci.* **2012**, *1*, 121–128, doi:10.1016/j.jshs.2012.07.007.
70. Denteneer, L.; Van Daele, U.; Truijen, S.; De Hertogh, W.; Meirte, J.; Stassijns, G. Reliability of physical functioning tests in patients with low back pain: A systematic review. *Spine J.* **2018**, *18*, 190–207, doi:10.1016/j.spinee.2017.08.257.
71. Denteneer, L.; Stassijns, G.; De Hertogh, W.; Truijen, S.; Van Daele, U. Inter- and Intrarater Reliability of Clinical Tests Associated With Functional Lumbar Segmental Instability and Motor Control Impairment in Patients With Low Back Pain: A Systematic Review. *Arch. Phys. Med. Rehabil.* **2017**, *98*, 151–164.e6, doi:10.1016/j.apmr.2016.07.020.
72. Nunnally, J.C.; Bernstein, I.H. The Assessment of Reliability. *Psychom. Theory* **1994**, *3*, 248–292.
73. Sanchez-Meca, J.; Rubio-Aparicio, M.; Marin-Martinez, F.; Lopez-Pina, J.A. The phenomenon of reliability induction in social and health sciences. In *Fifteenth Congress of Methodology of Behavioral Science*; Publisher: City, Spain, 2015.
74. Sánchez-Meca, J.; López-Pina, J.A.; López-López, J.A. Generalización de la fiabilidad: Un enfoque meta-analítico aplicado a la fiabilidad. *Fisioterapia* **2009**, *31*, 262–270, doi:10.1016/j.ft.2009.05.005.