Datson, N, Lolli, L, Drust, B, Atkinson, G, Weston, M and Gregson, W

Inter-methodological quantification of the target change for performance test outcomes relevant to elite female soccer players

Inter-methodological quantification of the target change for performance test outcomes relevant to elite female soccer players

Naomi Datson[1 2], Lorenzo Lolli[2], Barry Drust[3], Greg Atkinson[4], Matthew Weston[4] and Warren Gregson[2]

[1] Institute of Sport, University of Chichester, Chichester, UK
[2] Football Exchange, Research Institute of Sport Sciences, Liverpool John Moores University, Liverpool, UK
[3] School of Sport, Exercise and Rehabilitation Sciences, University of Birmingham, Birmingham, UK
[4] School of Health and Life Sciences, Teesside University, Middlesbrough, UK

Address for Correspondence:

Naomi Datson PhD
Institute of Sport
University of Chichester
Chichester
UK
N.Datson@chi.ac.uk

Abstract

Valid and informed interpretations of changes in physical performance test data are important within athletic development programmes. At present there is a lack of consensus regarding a suitable method for deeming whether a change in physical performance is practically-relevant or not. We compared true population variance in mean test scores between those derived from evidence synthesis of observational studies to those derived from practioner opinion (n=30), and to those derived from a measurement error (minimal detectable change) quantification (n=140). All these methods can help to obtain "target" change score values for performance variables. We found that the conventional "blanket" target change of 0.2 (between-subjects SD) systematically underestimated practically relevant and more informed changes derived for 5-m sprinting, 30-m sprinting, CMJ, and Yo-Yo Intermittent Recovery Level 1 (IR1) tests in elite female soccer players. For the first time in the field of sport and exercise sciences, we have illustrated the use of a principled approach for comparing different methods for the definition of changes in physical performance test variables that are practically-relevant. Our between-method comparison approach provides preliminary guidance for arriving at target change values that may be useful for research purposes and tracking of individual female soccer player's physical performance.

Key Words

Fitness testing, football, practically relevant change, player tracking, physical performance

Introduction

Physical performance testing is an integral component of an elite soccer player's development programme and is considered important by coaches, practitioners and players [1, 2]. Such performance assessments offer an opportunity to evaluate a player's physical qualities, and the derived information can be used to provide coaches and practitioners with evidence to guide talent identification, player selection and development programmes [3]. In the sports and exercise sciences, published research [4] has failed to provide information that may enable adequate study planning and facilitate meaningful interpretations of physical performance test data in the real-world [5]. It remains under-explored whether methods used to interpret *group-level* research [6] might be of any value to inform tracking processes at the *individual-level* [7-10]. We highlight that "*individual-level*" refers to individual-player data gathered in daily practice, whereas "*group-level*" indicates the aggregation of individual-player data for research purposes [7]. Real-world practice conventionally involves the examination and interpretation of individual-level (player) data [7].

Given that physical performance assessments are used to inform decision making throughout the player development process [11], robust interpretation of test performance data is, therefore, paramount [12]. In sports performance research, investigators are usually concerned with the determination of a group-level reference threshold, termed the smallest worthwhile change (SWC) or "target change", which is considered the 'practically relevant' change in the measure of interest [6]. In practice, changes in test score may be interpreted using the SWC statistic computed as *i*) percentage change or *ii*) some specified fraction of the available sample standard deviation (i.e., standardised effect size) [6]. However, the conceptual and contextual inconsistencies of these approaches limit the value of the SWC in the real-world [5, 13-19]. First, the calculation of pre-to-post changes expressed as percentage changes does not necessarily remove the regression-to-the-mean artefact that is a problem in single sample intervention or observational studies typical in this research field [17]. Second, use of standardised effect sizes (i.e., Cohen's *d*) to inform relevant interpretations can be misleading given the sample variance dependence and unitless expression lacking biological meaningfulness [15, 16, 18, 20]. Third, determining the importance of a change based on a magnitude scale as a fraction of a given sample standard deviation, generally 0.2 × between-subjects standard deviation (SD) [6], may be irrelevant in the context of sports performance [13, 18]. For example, a recent study on between-device measurement equivalence for maximal sprinting speed assessment showed how these criteria lack practical context [20]. Specifically, taking 0.2 × between-subjects SD as the target effect would have represented an unrealistic value for interpreting differences between the criterion and non-criterion measure considering what practitioners deemed meaningful [20].

There can be confusion over the different ways that target change thresholds are formulated and interpreted [21, 22], especially in terms of the distinction between minimal detectable change and minimal important change [21]. Minimal detectable change indicates the change in test performance beyond random within-subject variability of the measurement [23]. Conversely, minimal important change refers to the smallest change in a score domain of interest that players and coaches may perceive as meaningful [13, 24]. In practice, the minimal detectable change is based on a statistical threshold, whereas a minimal important change may be set irrespective of whether it can be distinguished from measurement error or not [25]. Likewise, the notion of practical relevance versus clinical relevance requires differentiation [22]. Practical relevance refers to whether the size of a change between two testing occasions can be said to differ reasonably [22]. Clinical relevance denotes whether the applied value of any observed change makes a real impact on overall sport performance from an empirical perspective [13, 22]. In general, tracking physical performance changes in the individual athlete is related to the notion of practical relevance.

Despite the current lack of consensus regarding established methods for specifying target change values [26-28], a general and perhaps arbitrary selection of a "global" target change may not necessarily coincide with a principled determination of a practically relevant change in performance variables on the actual scale of measurement [24, 29]. In the absence of objective information, the comparison of different methods involving data from existing sources of information and insight from practitioners can serve to provide guidance for real-world player tracking and research purposes [5, 13, 24, 30]. For example, the sports performance researcher may define the change values by comparing relevant information based on research evidence synthesis [31], distribution-based methods [32-34], and practitioner opinion [35, 36]. A systematic review and meta-analysis of observational data may be useful to inform the definition of a target change [30] that may be expressed as the population spread for the range of true mean population test scores. In line with its use in other fields of research [37], the tau-statistic is a standard deviation that indicates the variation across a *distribution* of true mean test scores [38] *beyond* random sampling error [39], and may be considered a relevant approximation for the definition of a practically relevant change of interest [40, 41]. The surveying of opinions from practitioners in the field also constitutes another valuable method for specifying change values deemed realistic as opposed to any potential guidance resting solely on statistical criteria [30, 35, 36]. Measurement error assessment is also important to understand whether a particular test may be useful for real-world practice [12, 42, 43]. Formal quantification of the minimal detectable change is relevant to ascertain whether any observed change can be distinguished from test-retest error [25, 44].

With information that can be obtained from systematic evidence synthesis, practitioner opinion and measurement error assessment, this study aimed to compare different methods for determining practically relevant changes in physical performance test variables relevant to elite female soccer players.

Materials and Methods

Systematic review and meta-analysis

Literature search procedures

Given the context of our study, we pre-determined relevant eligibility criteria [45] to inform our systematic review procedures (Table 1). A comprehensive electronic database search was conducted in PubMed and Web of Science by the lead author (ND) to identify original research articles published from the earliest record up to April 2020. A Boolean search phrase was created to include search terms relevant to the sport (soccer), sex (female) and physical performance test of interest (5-m sprinting, 30-m sprinting, CMJ), Yo-Yo IR1). Relevant keywords for each search term were determined through pilot searching (screening of titles, abstracts, keywords, and full texts of previously known articles). Keywords were combined within terms using the 'OR' operator, and the final search phrase was constructed by combining the three search terms using the 'AND' operator (Supplementary Table 1). All references were downloaded into a dedicated Papers library (Papers version 3.4.18). The library was reviewed, and duplicate records were identified and removed. After the removal of duplicate records, the title and abstracts of the remaining studies were screened against the inclusion and exclusion criteria (Table 1).

Data extraction

The full-text versions of the remaining articles were then retrieved and evaluated against the inclusion criteria to determine their final inclusion/exclusion status by the lead author (ND) and verified by one of the co-authors (LL). Full-text articles that met each of the eligibility criteria were included in quantitative synthesis, with a complete overview of the process for each test performance measure illustrated in Fig. 1-3. Consensus on study selection and data extraction was sought in meetings between the two reviewers throughout the process [46], with the sixth author (WG) consulted if necessary. Mean test scores and sampling variance were extracted by the lead author (ND) and subsequently verified by one of the co-authors (LL) for the observational studies meeting our eligibility criteria. Importantly, only baseline test performance measures were extracted in the case of experimental study designs, while a graph digitizer software (DigitizeIt, Braunschweig, Germany) facilitated the data extraction process where only scatter plots were available. The primary outcome to be reported from our evidence synthesis was the τ-statistic value [39, 47] as an approximation of the population standard deviation [48, 49] of true mean test scores.

Practically relevant changes in physical performance measures survey

Survey design and distribution

To obtain information relating to practically relevant changes in physical performance in female soccer, we conducted a short cross-sectional survey from July 2019 to April 2020. Practitioners (sport scientists, strength and conditioning coaches and fitness coaches) currently working in elite female soccer were asked on their perception of a practically relevant change in a range of physical performance tests (CMJ, 5-m and 30-m linear speed, and Yo-Yo IR1). The survey was developed in-house by the authors who represent a broad range of relevant expertise and experience in the area, both practically and scientifically [20]. The survey consisted of nine questions, covering two main areas: 1) introduction and background information (four questions), and 2) perceptions of change values across different physical performance tests (five questions). The data were collected using an online survey platform (Online Surveys, formerly Bristol Online Surveys). A weblink to the survey was generated and emailed with a covering letter to known contacts. The survey was intentionally distributed privately to known contacts to ensure completion by appropriate practitioners with the required experience within female soccer. Voluntary informed consent was requested at the start of the survey and no information regarding participant age, sex or club/national team was requested.

Measurement error assessment

Design

Physical performance tests were conducted on two separate occasions separated by seven days. All testing took place during the non-competitive phase of the season. Prior to assessment, all players had previously completed each test on at least one previous occasion, which acted as their familiarisation. All physical performance tests were performed on third generation turf (indoor arena) and players wore shorts, t-shirt and football boots (except for the jumps when trainers were worn). Players performed a standardised, generic warm-up prior to commencement of the physical assessments. All physical performance tests were completed at approximately the same time of day to reduce any circadian rhythm effect [50]. Tests were completed in a single session and in the same order (CMJ, linear speed and Yo-Yo IR1) on each test occasion. Test order was designed in an attempt to minimise the influence of previous tests on subsequent performance. Participants were instructed to refrain from strenuous exercise in the 24 hours before the fitness testing session and to consume their normal pre-training diet. To encourage maximal effort, players received consistent verbal encouragement throughout the physical performance tests. Overall, test-retest data were collected from 140 national team female soccer players (age range: 12 to 33 years). Usual appropriate ethics committee clearance was not required as data was collected as a condition of employment [49] and all players had previously consented for their data to be used for research purposes. Nevertheless, all data were anonymised prior to analysis to ensure player confidentiality.

Procedures

A standardised warm-up was completed, consisting of generic warm-up activity prior to commencing the physical performance tests. Specific warm-ups were also completed prior to each of the performance tests. To ensure consistency between testing occasions, National federation staff coached the warm-up activity.

Countermovement jump (without arms)

Estimations of player's lower limb muscular power were assessed via a countermovement jump (CMJ) on a jump mat (KMS Innervations, Australia). The jump mat was placed on a firm, concrete surface at the edge of the third-generation turf (indoor arena). Following the generic and jump-specific warm-up activity, the player was permitted an additional practice jump on the mat before performing three recorded trials. The player was instructed to step on to the mat and place their feet in the middle of the mat (a comfortable distance apart) and with their hands on their hips. The player started from an upright position and was instructed to jump as high as possible while keeping their hands on their hips. Players were instructed to keep their legs straight whilst in the air and refrain from bringing their legs into a pike position or flicking their heels. The highest jump height recorded to the nearest 0.1 cm was used as the criterion measure of performance.

Linear speed

Players' linear speed times were evaluated using electronic timing gates (Brower TC Timing System, USA) over distances of 0-30 m. A 50 m steel tape measure (Stanley, UK) was used to measure the 30 m distance and markers were placed at 0, 5 m and 30 m, in addition, a marker was placed 1 m behind the zero line. Tripods were placed directly over each marker at a height of 0.87 m above ground level and a timing gate (transmitter) was fitted to each tripod. Opposite each tripod, at a distance of 2 m, another tripod and timing gate (receiver) was positioned. Following the generic and speed-specific warm-up activity, the player was permitted an additional practice sprint through the course before performing three recorded trials. Each sprint was separated by a 3-min recovery period. The player commenced each sprint with their preferred foot on a line 1 m behind the first timing gate. The fastest time at each distance to the nearest 0.01 s was used as the criterion measure of performance.

Yo-Yo Intermittent Recovery Test Level 1

Estimations of player's high-intensity endurance capacity were assessed using the Yo-Yo Intermittent Recovery Test Level 1 (Yo-Yo IR1). During the test, participants completed a series of repeated 20 m shuttle runs with a progressively increasing running speed (10-19 $km \cdot h^{-1}$) interspersed with 10 s rest intervals [51].

Statistical analysis

Second-order information criterion (AICc) [52] assessed the relative quality of different models for meta-analysis with method of moments, maximum likelihood, and model error variance estimators for the true tau-statistic (τ) value [39]. By definition, the τ is a standard deviation describing the typical population variability across the distribution of true mean test scores given the summarised effects [39]. With different approaches described in the current literature [53], recent recommendations on methods for research evidence synthesis informed the meta-analytical framework of the present study [39, 47]. The methods selected to estimate the between-effect variance and its uncertainty involved the comparison of seven random-effects models using the DerSimonian-Laird, Hedges-Olkin, Sidik-Jonkman, maximum-likelihood, restricted maximum-likelihood, empirical Bayes, and Paule-Mandel estimators, respectively [39, 54]. The generalised Q-statistic method estimated the uncertainty around the mean τ-statistic value and was reported as 95% confidence interval (CI) [55]. The AICc difference (ΔAICc) from the estimated best model (i.e., the model with the lowest AICc value; ΔAICc = 0) was evaluated according to the following scale: 0-2, essentially equivalent; 2-7, plausible alternative; 7-14, weak support; > 14, no empirical support [56]. Results were interpreted from the best meta-analytical model for the examined data. Results from essentially equivalent models were also presented. Weighted raw point estimates were calculated as descriptive statistics with the 95% prediction interval (95% PI) describing the expected range for the distribution of true mean test scores for 95% of similar future studies [38, 57, 58]. All meta-analyses were performed using the *metafor* package [54].

Survey data were summarised as response frequency (expressed as counts or percentage) for categorical data, median and interquartile range (IQR) for count data and mean and standard deviation (SD) for continuous data. The change value in physical test performance measures practitioners deemed of practical relevance to elite female soccer was defined as mean and 95%CI from the available survey responses.

For the test-retest error assessment analyses, a paired samples t-test quantified the within-subjects SD for the mean difference in the test scores [12]. Random within-subject variability was quantified as the standard error of the measurement (SEM) [12] and presented with the respective uncertainty [59]. To assess absolute agreement between measurements [12], percentage coefficient of variation (%CV) was estimated using the logarithmic method [60, 61]. The minimal detectable change value for each performance measure was calculated as the product of the SEM value times 1.96 and the square root of 2 [42]. The underlying patterns in the raw test-retest data on each occasion were explored and illustrated in raincloud plots [62].

Effects for each selected method were presented and compared using density strips to illustrate the uncertainty (95%CI) surrounding the point estimates [63-65]. Statistical analyses were conducted using R (version 3.6.1, R Foundation for Statistical Computing).

Results

Systematic review and meta-analysis

Of the records we screened by title and abstract, 11, 17, 27, and 23 studies met the eligibility criteria for the 5-m sprinting [4, 66-75], 30-m sprinting [4, 76, 3, 77, 66, 67, 69, 78, 79, 72, 71, 80-84], CMJ [85-87, 3, 88, 76, 89, 69, 78, 90-92, 72, 93-97, 82, 73, 98-104], and Yo-Yo IR1 [105-108, 3, 76, 89, 109, 110, 69, 111, 78, 112, 71, 113, 114, 97, 99, 115-119] variables, respectively (Fig. 1-3). The identified samples of studies summarize almost twenty years of research on female soccer published between 2000 and 2020 encompassing test performance data ranging from youth to senior players. According to the model comparison on information-theory grounds (Supplementary Tables 2-5), the mean for the distribution of true mean test scores was 1.16 s (95%PI, 0.98 s to 1.34 s) for 5-m sprinting, 5.01 s (95%PI, 4.19 s to 5.83 s) for 30-m sprinting, 29 cm (95%PI, 21 cm to 37 cm) for CMJ, and 1077 m (95%PI, 527 m to 1628 m) for Yo-Yo IR1.

Practically relevant changes in physical performance measures survey

Median time (IQR) to complete the survey (min:sec) was 08:31 (03:29 to 19:57). Of the 30 respondents, 63% were strength and conditioning coaches and 30% sports scientists (Q1). Respondents had a median of 3 (2 to 6) years of experience working in female soccer (Q2), and worked either in senior (37%), youth (30%), or combination of both (33%) female soccer contexts at the time surveyed (Q3). The majority of respondents worked with National teams or clubs in the top division in their respective country (73%) (Q4), with the following breakdown of leagues/level of competition that respondents clubs played in: National teams (n = 8), English Women's Super League (n = 6), English Women's Championship (n = 3), Italian Serie A (n = 3), Australian W League (n = 2), English Regional Talent Club (n = 2), English National Premier League (n = 1), USA National Women's Soccer League (n = 1), USA National Collegiate Athletic Association (n = 1), French Division 1 Feminine (n = 1), Northern Ireland Women's Premiership (n = 1), and highest league (country not stated) (n = 1).

Measurement error assessment

The estimated mean test-retest difference was 0.002 s (95%CI, −0.004 s to 0.007 s), −0.015 s (95%CI, −0.029 s to −0.002 s), 0.01 cm (95%CI, −0.24 cm to 0.26 cm), and −16 m (95%CI, −33 m to 2 m) for 5-m sprinting, 30-m sprinting, CMJ, and Yo-Yo IR1 variables, respectively. The %CV (95%CI) was 2.3% (2.0% to 2.6%) for 5-m sprinting, 1.2% (1.1% to 1.4%) for 30-m sprinting, 3.9% (3.4% to 4.3%) for CMJ, and 7.2% (6.3% to 8.1%) for Yo-Yo IR1 data. Raincloud plots illustrated the data distribution and degree of raw test-retest measurement error (Fig. 4).

Between-method comparison

5-m sprinting

Formal comparison of different meta-analytical approaches revealed the random-effects model with maximum likelihood estimator for the τ to be the best of the seven candidates (Supplementary Table 2). The τ was ± 0.08 s (95%CI, 0.06 s to 0.14 s). All the essentially equivalent models provided similar values for the point estimate based on a sample of 272 female players. Given the observed degree of test-retest measurement error (Fig. 4), the calculated minimal detectable change value in 5-m sprinting performance was ± 0.07 s (95%CI, 0.06 s to 0.08 s). The survey results suggested a mean change of ± 0.09 s (95%CI, 0.04 s to 0.13 s). In contrast, use of the "test" reliability data for the calculation of small effect in Cohen's terms (0.2 × between-subjects SD) underestimated the change value (Δ =± 0.011 s; 95%CI, 0.010 s to 0.012 s).

30-m sprinting

The random-effects model with maximum likelihood estimation method for the τ was the best in the pool of candidates (Supplementary Table 3). Meta-analyses involved 685 female players revealed a τ value of ± 0.39 s (95%CI, 0.31 s to 0.57 s), with essentially equivalent models providing similar estimates. The calculated minimal detectable change value was ± 0.16 s (95%CI, 0.14 s to 0.18 s) on the basis of the test-retest measurement error analyses (Fig. 4). The mean change practitioners perceived as practically relevant was ± 0.21 s (95%CI, 0.11 s to 0.32 s). Estimation of a small effect as per Cohen's criteria using "test" reliability data yielded an underestimated change value of ± 0.044 s (95%CI, 0.040 s to 0.050 s).

CMJ

Following our meta-analytical model comparison on information-theory grounds, the random-effects model with maximum likelihood estimator was found to be the best relative to other competing models (Supplementary Table 4). With an available dataset including 1792 female players, the estimated τ was ± 3.9 cm (95%CI, 3.3 cm to 4.9 cm). The estimated minimal detectable change value was ± 2.9 cm (95%CI, 2.6 cm to 3.3 cm), while the mean change value perceived as important by practitioners was ± 2.8 cm (95%CI, 2.1 cm to 3.4 cm). The change value of ± 1.0 cm (95%CI, 0.9 cm to 1.1 cm) commensurate to a small effect according to Cohen was inconsistent with the all the mean estimates obtained from the other approaches.

Yo-Yo IR1

The AICc criteria revealed the random-effects model with restricted maximum likelihood estimator for the τ as the best model in the set of candidates (Supplementary Table 5). Using available Yo-Yo IR1 data from an overall sample of 981 female players, the τ was ± 267 m (95%CI, 210 m to 355 m). Given the observed random-within subject variability in the Yo-Yo IR1 assessment, the calculated value for the minimal detectable change was ± 206 m (95%CI, 184 m to 233 m). The mean value for the change deemed of practical relevance by practitioners was ± 164 m (95%CI, 123 m to 206 m). Conversely, use of the "test" reliability data for calculation of the change as per Cohen's criteria (0.2 × between-subjects SD) yielded an underestimated value of ± 92 m (95%CI, 82 m to 104 m).

Discussion

Using a principled approach in the domain of sport and exercise sciences, this is the first study to illustrate a formal comparison of different methods for determining practically relevant target change values in physical performance test variables. Our study findings suggested that the definition of a target change value depends on the context and purpose of the measurement.

Despite the lack of consensus regarding a standardized methodology for defining change values [26, 27], an a priori and arbitrary selection of a single method is unlikely to result in a rationalised determination of practically relevant changes on the actual scale of measurement [24, 34]. Establishing a change value of interest has inherent challenges, but is considered relatively straightforward in sports such as cycling or running, whereby the performance outcome is usually time or distance [13, 24]. Conversely, determining a practically relevant change in a multi-component sport such as soccer or rugby is more challenging and thus consideration of between-method comparisons appears relevant irrespective of the context [41]. Specifically, the degree of a target change may differ if considered from research and applied perspectives and not correspond to a fixed or universal value that may be of interest to different stakeholders [8]. Values deemed meaningful for group-level research may not be applicable for individual-player tracking purposes [120]. The sports performance researcher would consider a target change to inform study design, while the practitioner is concerned with changes which guide player evaluation strategies [8]. The general strategy of inter-methodological quantification of target changes intends to stimulate further discussion between the researcher and practitioner, not an end in itself. For example, adequate sample size planning requires explicit specification of an effect of interest [30], yet researchers typically rely on unjustified conventions not calibrated to any study context [121]. Failure to specify what change would falsify a research hypothesis may lead to unnecessarily inconclusive studies and ambiguous interpretations of findings [30, 122]. Use of information from practitioner opinion (i.e., opinion-seeking method) would be preferable if one aims to assess whether an intervention elicited within-individual changes greater than change values deemed realistic and relevant to interpretation of research findings (i.e., group-level research) [36, 123]. The choice of this or any alternative method for player tracking purposes would, however, depend on whether one is interested in evaluating the size or the meaning of a change for overall sports performance [13].

Measurement error assessment can represent a first step to support interpretations when no empirical guidance is available and should be complementary to other methods [44, 124]. This particular evaluation is only useful for understanding whether a change value can be distinguished from random within-subject variability [124]. Measurement reliability should not constitute a proxy for determining what value may be judged practically or clinically relevant [25]. However, a practically relevant change smaller than a minimal detectable change may not be distinguished from measurement error irrespective of the purpose. Research in clinimetrics highlighted the importance of reducing measurement error, not increasing the value of a target change [124]. In practice, if a change deemed relevant by practitioners equals 1 standard error of the measurement, the minimal detectable change will always be systematically larger [124]. In our study, the use of test-retest data from 140 national team female soccer players (age range: 12 to 33 years) enabled an estimation of the error in each performance test free from the influence of sampling imprecision. The fact that the mean target change for the Yo-Yo IR1 performance test based on practitioner opinion did not exceed the measurement error value (Figure 5) suggested it may not be helpful for tracking high-intensity endurance performance in the individual player [9]. To illustrate this from a practical perspective, the derived change for Yo-Yo IR1 performance from each approach was; ± 267 m (evidence synthesis), ± 206 m (test-retest measurement error assessment) and ± 164 m (practitioner opinion). In contrast, change values derived from practioners' opinions and alternative distribution-based methods were larger than measurement error-based values for interpretations of data relevant to sprint and jump variables. Our study confirmed that changes deemed practically relevant by practitioners may not converge to a consistent range of values determined by the error of the measurement scale or other distribution-based criteria for each performance variable of interest. Any decision for selecting one or another value informed by, for example, the range of target changes we described as in the case of the Yo-Yo IR1 variable should be pragmatic and based on the context of the measurement [8, 120].

In the sport and exercise sciences, the general practice among researchers and practitioners typically involves the derivation of practically relevant changes as a function of arbitrary fractions of one-off sample standard deviation by calculating the value of interest as 0.2 × between-subjects SD of the previous assessment data [6]. The sample-dependent nature of this approach is a major drawback precluding the definition of changes having relevance for research and real-world practice. Formal comparison of results from different methods indicated that determination of a change score as a *small* effect according to Cohen's criteria [125] systematically underestimated the value of interest when compared to the other approaches considered in this study. In this context, a recent study illustrated the discrepancy between the use of these criteria and more rationalised methods as practitioner opinion to arrive at values deemed realistic [20]. As a consequence, practitioners should be wary of interpreting changes in performance assessments based on the conventional 0.2 × between-subjects SD criterion a priori [6]. Our preliminary findings were in line with recent observations discouraging any specious reliance on effect sizes as limited measures of practical relevance [18, 19, 126].

The available information in this and other research fields guided the selection of different methods to address specific aspects in our study [24, 25, 33, 40, 123]. As a distribution-based method, consideration of the variation in a group of test scores is a typical approach used to inform the definition of practically relevant effects [40]. Norman and colleagues emphasised how change values defined on statistical criteria from individual studies per se might depend unnecessarily on sampling and inherent characteristics [41]. Accordingly, the synthesis of observational data illustrated in this study aimed to describe an approximation of a population variation value for each test measure [48, 49] that may be realistic and generalisable beyond the single study of limited size [127]. Quantifying the amount of change needed to be certain that a given change that occurred was beyond measurement error is another criterion generally adopted by clinical researchers [123]. Acknowledging the fundamental distinction between statistical and principled criteria [25], the minimal detectable change may be an informative benchmark when no empirical guidance is available as in our study context. Nevertheless, the basis of any estimate derived from these or any other plausible approach rests on a formal appraisal of their potential importance [123]. Opinion-seeking represents a method valuable for maximising the practical context of findings to assess expectations regarding what is deemed realistic by practitioners [30]. In this respect, findings from this method can represent a critical counterpoint to what might be viewed achievable solely on statistical grounds. Nevertheless, in practice, how it should be weighed compared to other methods remains unexplored.

The process for the definition of practically relevant changes in physical performance measures may also require careful considerations inherent to the application of group-based values for the screening of the individual player [7, 128-131] and the presence of other available alternatives, as, for example, anchor-based approaches. Adoption of this method involves the comparison of a player's test performance on two different occasions and then relating the observed change score to a predetermined, independent measure or “anchor" [26, 33, 132]. The anchor is interpretable itself (e.g., self-reported outcome measures on a psychometric scale) and, for example, can be based either on player, coach or practitioner judgements of perceived improvement or deterioration in test performance on a given assessment [123, 133]. Nevertheless, it is important to emphasise that the practical value of determining change values using anchor-based methods relies on a well-conceived study design [133, 134]. The extent of anchor-based estimates is dependent on the selection of the anchor itself, which may vary substantially between different perspectives and contexts [5, 13, 29, 28, 123]. In this, and other fields of research, there is no empirical guideline on how and whether the application of group-based results (between-subjects approach) from sports science studies may be valid to inform the monitoring of the individual player over time (within-subjects approach) [28]. Beaton et al., [130] maintained that the magnitude of a change value could substantially differ when comparing between-subjects versus within-subjects methods considering these as conceptually different approaches. Cella et al., [128] however, argued that group-derived data can be used to inform the interpretation of changes at the individual-subject level, but not without the support of relevant information inherent to random within-subject variability. What emerged from our comparison of between-subjects (e.g., meta-analysis) and within-subjects (e.g., practioner opinion and measurement error assessment) approaches suggested methods should be seen as complementary to each other to arrive at rationalised interpretations of measurements in research and real-world practice [135].

Our study is not without limitations. Our investigation did not provide information regarding our survey content validity since the instrument did not undergo a formal pilot phase. However, we did not consider that as necessary due to the fundamental simplicity of our survey. As illustrated in a recent study [20], our survey focused primarily on one question regarding practitioners' perspectives on change values perceived as meaningful and relevant to the interpretation of different physical performance test scores. Specifically, the notion of meaningful referred to the degree of an observed change on that particular test and not its relative contribution to a potential enhancement in overall soccer performance [13]. The synthesis of observational data derived in independent groups both in different studies and within the same study is another aspect to consider [136]. Also, our selection [123] of some among other potential methods for specifying a change value of interest requires careful consideration. The relevance of available methods arguably depends on the research aim and context [8, 40]. Clinical researchers highlighted both values and limitations of using distribution-based methods, opinion-seeking, and review of the evidence base for specifying an effect deemed of *minimal importance* [18, 24, 28, 34, 40, 123, 137]. Likewise, taking into consideration the initial test performance level can be important for the definition and interpretation of a practically relevant change in the measure of interest [33]. Consideration of the initial test performance level assumes that greater changes between testing occasions for subjects with lower initial performance are the consequence of functional adaptations only [33]. However, this tendency may just be as consistent with the effects of the regression-to-the-mean artifact whereby more extreme scores can become less extreme at a follow-up assessment [33]. In practice, subjects with relatively higher test scores will find it harder to attain a given change when compared to subjects with relatively lower test scores [33]. Accounting for this important aspect may limit arriving at conclusions that subjects with relatively lower test scores attained true practically relevant changes in test performance [33]. Different approaches were applied in the clinical literature [33] and recently in the sports sciences [138], although there is no consensus on an established method to address this particular statistical phenomenon. Likewise, accounting for the player's perspective on changes in test scores and performance outcomes beyond opinion-based or statistical criteria would be of great importance [128, 139]. Given our data, exploration of these particular aspects was not, however, practically feasible thereby suggesting caution when generalizing what is illustrated in the present study.

Conclusion

This study compared different methods for defining practically relevant changes in physical performance measures. Our results highlighted how information obtained from between-method comparisons could be superior to *any a priori* adoption of conventional statistical criteria (e.g., 0.2 × between-subject SD) to support more rationalised interpretations of individual player test scores and research findings. The specification of a target change in physical performance tests is context-specific and should not be determined *a priori* on one study or one method only. Our findings provide guidance that may be useful for research purposes and tracking the physical performance of individual elite female soccer players in the absence of more objective information.

Acknowledgments

The authors would like to express their gratitude to the English FA for providing access to the current data as well as staff and players for their co-operation during data collection. The authors would also like to recognise and thank the practitioners who kindly completed the survey.

Disclosure of Interest

The authors report no conflict of interest

References

1. Hulse MA, Morris JG, Hawkins RD, Hodson A, Nevill AM, Nevill ME. A field-test battery for elite, young soccer players. Int J Sports Med. 2013;34(4):302-11. doi:10.1055/s-0032-1312603.
2. Manson SA, Brughelli M, Harris NK. Physiological characteristics of international female soccer players. J Strength Cond Res. 2014;28(2):308-18. doi:10.1519/JSC.0b013e31829b56b1.
3. Datson N, Weston M, Drust B, Gregson W, Lolli L. High-intensity endurance capacity assessment as a tool for talent identification in elite youth female soccer. J Sports Sci. 2019:1-7. doi:10.1080/02640414.2019.1656323.
4. Datson N, Hulton A, Andersson H, Lewis T, Weston M, Drust B et al. Applied physiology of female soccer: an update. Sports Med. 2014;44(9):1225-40. doi:10.1007/s40279-014-0199-1.
5. Atkinson G. What’s behind the numbers? Important decisions in judging practical significance. Sportscience. 2007;11:12-5.
6. Buchheit M. The numbers will love you back in return-I promise. Int J Sports Physiol Perform. 2016;11(4):551-4. doi:10.1123/ijspp.2016-0214.
7. King MT, Dueck AC, Revicki DA. Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? Med Care. 2019;57 Suppl 5 Suppl 1:S38-s45. doi:10.1097/mlr.0000000000001111.
8. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol. 2002;14(2):109-14. doi:10.1097/00002281-200203000-00006.
9. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V et al. Minimal clinically important differences: review of methods. J Rheumatol. 2001;28(2):406-12.
10. Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. Med Care. 2003;41(5):593-6. doi:10.1097/01.Mlr.0000064706.35861.B4.
11. Svensson M, Drust B. Testing soccer players. J Sports Sci. 2005;23(6):601-18. doi:10.1080/02640410400021294.
12. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998;26(4):217-38.
13. Atkinson G. Does size matter for sports performance researchers? J Sports Sci. 2003;21(2):73-4. doi:10.1080/0264041031000071038.
14. Bland JM. The tyranny of power: is there a better way to calculate sample size? BMJ. 2009;339:b3985. doi:10.1136/bmj.b3985.
15. Lenth R. Some practical guidelines for effective sample size. Am Stat. 2001;55(3):187-93. doi:10.1198/000313001317098149.
16. Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017;355(6325):584-5. doi:10.1126/science.aal3618.

17. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. BMJ. 2003;326(7398):1083-4. doi:10.1136/bmj.326.7398.1083.
18. Pogrow S. How effect size (practical significance) misleads clinical practice: the case for switching to practical benefit to assess applied research findings. Am Stat. 2019;73:223-34. doi:10.1080/00031305.2018.1549101.
19. Gibbs NM, Weightman WM. Beyond effect size: consideration of the minimum effect size of interest in anesthesia trials. Anesth Analg. 2012;114(2):471-5. doi:10.1213/ANE.0b013e31823d2ab7.
20. Kyprianou E, Lolli L, Al Haddad H, Di Salvo V, Varley M, Mendez-Villanueva A et al. A novel approach to assessing validity in sports performance research: integrating expert practitioner opinion into the statistical analysis. Sci Med Footb. 2019;3(4):333-8. doi:10.1080/24733938.2019.1617433.
21. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006;4:54. doi:10.1186/1477-7525-4-54.
22. Bothe AK, Richardson JD. Statistical, practical, clinical, and personal significance: definitions and applications in speech-language pathology. Am J Speech Lang Pathol. 2011;20(3):233-42. doi:10.1044/1058-0360(2011/10-0034).
23. Lassere MN, van der Heijde D, Johnson KR. Foundations of the minimal clinically important difference for imaging. J Rheumatol. 2001;28(4):890-1.
24. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102-9. doi:10.1016/j.jclinepi.2007.03.012.
25. de Vet HC, Terwee CB. The minimal detectable change should not replace the minimal important difference. J Clin Epidemiol. 2010;63(7):804-5; author reply 6. doi:10.1016/j.jclinepi.2009.12.015.
26. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol. 2010;63(5):524-34. doi:10.1016/j.jclinepi.2009.08.010.
27. Wright JG. The minimal important difference: who's to say what is important? J Clin Epidemiol. 1996;49(11):1221-2. doi:10.1016/s0895-4356(96)00207-7.
28. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res. 2011;11(2):171-84. doi:10.1586/erp.11.9.
29. Thorpe RT, Atkinson G, Drust B, Gregson W. Monitoring fatigue status in elite team-sport athletes: implications for practice. Int J Sports Physiol Perform. 2017;12(Suppl 2):S227-s34. doi:10.1123/ijspp.2016-0434.
30. Cook JA, Julious SA, Sones W, Hampson LV, Hewitt C, Berlin JA et al. DELTA(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. BMJ. 2018;363:k3750. doi:10.1136/bmj.k3750.
31. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-76. doi:10.1038/nrn3475.
32. Beaton DE, van Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA et al. Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. J Clin Epidemiol. 2011;64(5):487-96. doi:10.1016/j.jclinepi.2010.07.012.
33. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003;56(5):395-407. doi:10.1016/s0895-4356(03)00044-1.

34. Crosby RD, Kolotkin RL, Williams GR. An integrated method to determine meaningful changes in health-related quality of life. J Clin Epidemiol. 2004;57(11):1153-60. doi:10.1016/j.jclinepi.2004.04.004.
35. Staunton H, Willgoss T, Nelsen L, Burbridge C, Sully K, Rofail D et al. An overview of using qualitative techniques to explore and define estimates of clinically important change on clinical outcome assessments. J Patient Rep Outcomes. 2019;3(1):16. doi:10.1186/s41687-019-0100-y.
36. Fayers PM, Cuschieri A, Fielding J, Craven J, Uscinska B, Freedman LS. Sample size calculation for clinical trials: the impact of clinician beliefs. Br J Cancer. 2000;82(1):213-9. doi:10.1054/bjoc.1999.0902.
37. Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS et al. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004;57(9):898-910. doi:10.1016/j.jclinepi.2004.01.012.
38. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Method. 2010;1(2):97-111. doi:10.1002/jrsm.12.
39. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods. 2016;7(1):55-79. doi:10.1002/jrsm.1164.
40. Copay AG, Subach BR, Glassman SD, Polly DW, Jr., Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J. 2007;7(5):541-6. doi:10.1016/j.spinee.2007.01.008.
41. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Medical care. 2003;41(5):582-92. doi:10.1097/01.Mlr.0000062554.74615.4c.
42. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231-40. doi:10.1519/15184.1.
43. Hebert R, Spiegelhalter DJ, Brayne C. Setting the minimal metrically detectable change on disability rating scales. Arch Phys Med Rehabil. 1997;78(12):1305-8. doi:10.1016/s0003-9993(97)90301-4.
44. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol. 2009;62(10):1062-7. doi:10.1016/j.jclinepi.2008.10.011.
45. McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnston RV, Thomas J. Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane, 2020. Available from www.training.cochrane.org/handbook. 2020.
46. Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second reviewer for study selection in systematic reviews. Res Synth Methods. 2019;10(4):539-45. doi:10.1002/jrsm.1369.
47. Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods. 2019;10(1):83-98. doi:10.1002/jrsm.1316.
48. Altman DG. Practical statistics for medical research. London: Chapman and Hall/CRC; 1991.
49. Healy MJ. Populations and samples. Arch Dis Child. 1991;66(11):1355-6. doi:10.1136/adc.66.11.1355.

50. Reilly T, Brooks GA. Exercise and the circadian variation in body temperature measures. International journal of sports medicine. 1986;7(6):358-62. doi:10.1055/s-2008-1025792.
51. Krustrup P, Mohr M, Amstrup T, Rysgaard T, Johansen J, Steensberg A et al. The yo-yo intermittent recovery test: physiological response, reliability, and validity. Med Sci Sports Exerc. 2003;35(4):697-705. doi:10.1249/01.MSS.0000058441.94520.32.
52. Hurvich CM, Tsai CL. Regression and time-series model selection in small samples. Biometrika. 1989;76(2):297-307. doi:10.1093/biomet/76.2.297.
53. Petropoulou M, Mavridis D. A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: a simulation study. Stat Med. 2017;36(27):4266-80. doi:10.1002/sim.7431.
54. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):1-48.
55. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. Stat Med. 2007;26(1):37-52. doi:10.1002/sim.2514.
56. Burnham KP, Anderson DR, Huyvaert KP. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol. 2011;65(1):23-35. doi:10.1007/s00265-010-1029-6.
57. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172:137-59. doi:10.1111/j.1467-985X.2008.00552.x.
58. IntHout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. 2016;6(7):e010247. doi:10.1136/bmjopen-2015-010247.
59. Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. Chapman and Hall/CRC; 2000.
60. Bland JM, Altman DG. Measurement error proportional to the mean. BMJ. 1996;313(7049):106. doi:10.1136/bmj.313.7049.106.
61. Bland JM. How should I calculate a within-subject coefficient of variation? 2006. https://www-users.york.ac.uk/~mb55/meas/cv.htm.
62. Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome open research. 2019;4:63. doi:10.12688/wellcomeopenres.15191.1.
63. Bowman AW. Graphics for uncertainty. J R Stat Soc Ser A Stat Soc. 2019;182:403-18. doi:10.1111/rssa.12379.
64. Jackson CH. Displaying uncertainty with shading. Am Stat. 2008;62(4):340-7. doi:10.1198/000313008X370843.
65. Moore DS, McCabe GP, Craig BA. Introduction to the practice of statistics. W.H. Freeman and Company; 2007.
66. Baumgart C, Freiwald J, Hoppe MW. Sprint mechanical properties of female and different aged male top-level german soccer players. Sports (Basel, Switzerland). 2018;6(4). doi:10.3390/sports6040161.
67. Bishop C, Read P, McCubbine J, Turner A. Vertical and horizontal asymmetries are related to slower sprinting and jump performance in elite youth female soccer players. J Strength Cond Res. 2018. doi:10.1519/jsc.0000000000002544.
68. Gabbett TJ, Carius J, Mulvey M. Does improved decision-making ability reduce the physiological demands of game-based activities in field sport athletes? J Strength Cond Res. 2008;22(6):2027-35. doi:10.1519/JSC.0b013e3181887f34.
69. Hammami MA, Ben Klifa W, Ben Ayed K, Mekni R, Saeidi A, Jan J et al. Physical performances and anthropometric characteristics of young elite North-African female soccer

players compared with international standards. Science&Sports. 2020;35(2):67-74. doi:10.1016/j.scispo.2019.06.005.

70. Hoare DG, Warr CR. Talent identification and women's soccer: an Australian experience. J Sports Sci. 2000;18(9):751-8. doi:10.1080/02640410050120122.

71. Lockie RG, Moreno MR, Lazar A, Orjalo AJ, Giuliano DV, Risso FG et al. The physical and athletic performance characteristics of division I collegiate female soccer players by position. J Strength Cond Res. 2018;32(2):334-43. doi:10.1519/jsc.0000000000001561.

72. Julian R, Hecksteden A, Fullagar HH, Meyer T. The effects of menstrual cycle phase on physical performance in female soccer players. PLoS One. 2017;12(3):e0173951. doi:10.1371/journal.pone.0173951.

73. Pedersen S, Heitmann KA, Sagelv EH, Johansen D, Pettersen SA. Improved maximal strength is not associated with improvements in sprint time or jump height in high-level female football players: a clusterrendomized controlled trial. BMC Sports Sci Med Rehabil. 2019;11:20. doi:10.1186/s13102-019-0133-9.

74. Sport AIo. Physiological tests for elite athletes. 2nd ed. Champaign, United States: Human Kinetics Publishers; 2012.

75. Taylor JM, Portas M, Wright MD, Hurst C, Weston M. Within-season variation of fitness in elite youth female soccer players. J Athl Enhancement. 2012;2(1). doi:10.4172/2324-9080.1000102.

76. Emmonds S, Till K, J. R, Murray E, Turner L, Robinson C et al. Influence of age on the anthropometric and performance characteristics of high-level youth female soccer players. Int J Sports Sci Coa. 2018;13(5):779-86. doi:10.1177/1747954118757437.

77. Andersen E, Lockie RG, Dawes JJ. Relationship of absolute and relative lower-body strength to predictors of athletic performance in collegiate women soccer players. Sports (Basel, Switzerland). 2018;6(4). doi:10.3390/sports6040106.

78. Idrizovic K. Physical and anthropometric profiles of elite female soccer players. Med Sport. 2014;67(2):273-87.

79. Jackman SR, Scott S, Randers MB, Orntoft C, Blackwell J, Zar A et al. Musculoskeletal health profile for elite female footballers versus untrained young women before and after 16 weeks of football training. J Sports Sci. 2013;31(13):1468-74. doi:10.1080/02640414.2013.796066.

80. McFarland IT, Dawes JJ, Elder CL, Lockie RG. Relationship of Two Vertical Jumping Tests to Sprint and Change of Direction Speed among Male and Female Collegiate Soccer Players. Sports (Basel, Switzerland). 2016;4(1). doi:10.3390/sports4010011.

81. Nebil G, Zouhair F, Hatem B, Hamza M, Zouhair T, Roy S et al. Effect of optimal cycling repeated-sprint combined with classical training on peak leg power in female soccer players. Isokinet Exerc Sci 2014;22(1):69-76. doi:10.3233/IES-130515.

82. Oberacker LM, Davis SE, Haff GG, Witmer CA, Moir GL. The Yo-Yo IR2 test: physiological response, reliability, and application to elite soccer. J Strength Cond Res. 2012;26(10):2734-40. doi:10.1519/JSC.0b013e318242a32a.

83. Ozbar N. Effects of plyometric Training on explosive strength, speed and kicking speed in female soccer players. Anthropol. 2015;19(2):333-9. doi:10.1080/09720073.2015.11891666.

84. Ünveren A. Investigating women futsal and soccer players’ acceleration, speed and agility features. Anthropol. 2015;21(1-2):361-5.

85. Andersson H, Raastad T, Nilsson J, Paulsen G, Garthe I, Kadi F. Neuromuscular fatigue and recovery in elite female soccer: effects of active recovery. Med Sci Sports Exerc. 2008;40(2):372-80. doi:10.1249/mss.0b013e31815b8497.

86. Brannstrom A, Yu JG, Jonsson P, Akerfeldt T, Stridsberg M, Svensson M. Vitamin D in relation to bone health and muscle function in young female soccer players. Eur J Sport Sci. 2017;17(2):249-56. doi:10.1080/17461391.2016.1225823.
87. Castagna C, Castellini E. Vertical jump performance in Italian male and female national team soccer players. Journal of strength and conditioning research / National Strength & Conditioning Association. 2013;27(4):1156-61. doi:10.1519/JSC.0b013e3182610999.
88. Emmonds S, Nicholson G, Begg C, Jones B, Bissas A. Importance of physical qualities for speed and change of direction ability in elite female soccer players. J Strength Cond Res. 2019;33(6):1669-77. doi:10.1519/jsc.0000000000002114.
89. Francescato MP, Venuto I, Buoite A, Stel G, Mallardi F, Cauci S. Sex differences in hydration status among adolescent elite soccer players. J Hum. 2019;14(2):265-80. doi:10.14198/jhse.2019.142.02.
90. Haugen TA, Tonnessen E, Seiler S. Speed and countermovement-jump characteristics of elite female soccer players, 1995-2010. Int J Sports Physiol Perform. 2012;7(4):340-9. doi:10.1123/ijspp.7.4.340.
91. Ingebrigtsen J, Shalfawi SA, Tonnessen E, Krustrup P, Holtermann A. Performance effects of 6 weeks of aerobic production training in junior elite soccer players. J Strength Cond Res. 2013;27(7):1861-7. doi:10.1519/JSC.0b013e31827647bd.
92. Jeras NMJ, Bovend'Eerdt TJH, McCrum C. Biomechanical mechanisms of jumping performance in youth elite female soccer players. J Sports Sci. 2019:1-7. doi:10.1080/02640414.2019.1674526.
93. Krustrup P, Zebis M, Jensen JM, Mohr M. Game-induced fatigue patterns in elite female soccer. J Strength Cond Res. 2010;24(2):437-41. doi:10.1519/JSC.0b013e3181c09b79.
94. Lesinski M, Muehlbauer T, Granacher U. Concurrent validity of the Gyko inertial sensor system for the assessment of vertical jump height in female sub-elite youth soccer players. BMC Sports Sci Med Rehabil. 2016;8:35. doi:10.1186/s13102-016-0061-x.
95. Loturco I, Suchomel T, James LP, Bishop C, Abad CCC, Pereira LA et al. Selective Influences of Maximum Dynamic Strength and Bar-Power Output on Team Sports Performance: A Comprehensive Study of Four Different Disciplines. Frontiers in physiology. 2018;9:1820. doi:10.3389/fphys.2018.01820.
96. McCurdy KW, Walker JL, Langford GA, Kutz MR, Guerrero JM, McMillan J. The relationship between kinematic determinants of jump and sprint performance in division I women soccer players. Journal of strength and conditioning research / National Strength & Conditioning Association. 2010;24(12):3200-8. doi:10.1519/JSC.0b013e3181fb3f94.
97. Mujika I, Santisteban J, Impellizzeri FM, Castagna C. Fitness determinants of success in men's and women's football. J Sports Sci. 2009;27(2):107-14. doi:10.1080/02640410802428071.
98. Prieske O, Maffiuletti NA, Granacher U. Postactivation potentiation of the plantar flexors does not directly translate to jump performance in female elite young soccer players. Frontiers in physiology. 2018;9:276. doi:10.3389/fphys.2018.00276.
99. Ramos GP, Nakamura FY, Penna EM, Mendes TT, Mahseredjian F, Lima AM et al. Comparison of physical fitness and anthropometrical profiles among brazilian female soccer national teams from U15 to senior categories. Journal of strength and conditioning research / National Strength & Conditioning Association. 2019. doi:10.1519/jsc.0000000000003140.
100. Sedano S, Vaeyens R, Philippaerts RM, Redondo JC, Cuadrado G. Anthropometric and anaerobic fitness profile of elite and non-elite female soccer players. J Sports Med Phys Fitness. 2009;49(4):387-94.
101. Shalfawi SA, Haugen T, Jakobsen TA, Enoksen E, Tonnessen E. The effect of combined resisted agility and repeated sprint training vs. strength training on female elite

soccer players. J Strength Cond Res. 2013;27(11):2966-72. doi:10.1519/JSC.0b013e31828c2889.

102. Steffen K, Bakka HM, Myklebust G, Bahr R. Performance aspects of an injury prevention program: a ten-week intervention in adolescent female football players. Scand J Med Sci Sports. 2008;18(5):596-604. doi:10.1111/j.1600-0838.2007.00708.x.

103. Suchomel TJ, Sole CJ, Bailey CA, Grazer JL, Beckham GK. A comparison of reactive strength index-modified between six U.S. Collegiate athletic teams. J Strength Cond Res. 2015;29(5):1310-6. doi:10.1519/jsc.0000000000000761.

104. Vescovi JD, Rupf R, Brown TD, Marques MC. Physical performance characteristics of high-level female soccer players 12-21 years of age. Scand J Med Sci Sports. 2011;21(5):670-8. doi:10.1111/j.1600-0838.2009.01081.x.

105. Andersen TB, Krustrup P, Bendiksen M, Orntoft CO, Randers MB, Pettersen SA. Kicking velocity and effect on match performance when using a smaller, lighter ball in women's football. International journal of sports medicine. 2016;37(12):966-72. doi:10.1055/s-0042-109542.

106. Bendiksen M, Pettersen SA, Ingebrigtsen J, Randers MB, Brito J, Mohr M et al. Application of the Copenhagen soccer test in high-level women players - locomotor activities, physiological response and sprint performance. Hum Mov Sci. 2013;32(6):1430-42. doi:10.1016/j.humov.2013.07.011.

107. Booysen MJ, Gradidge PJ, Constantinou D. Anthropometric and motor characteristics of South African national level female soccer players. Journal of human kinetics. 2019;66:121-9. doi:10.1515/hukin-2017-0189.

108. Cone JR, Berry NT, Goldfarb AH, Henson RA, Schmitz RJ, Wideman L et al. Effects of an individualized soccer match simulation on vertical stiffness and impedance. J Strength Cond Res. 2012;26(8):2027-36. doi:10.1519/JSC.0b013e31823a4076.

109. Flatt AA, Esco MR. Evaluating individual training adaptation with smartphone-derived heart rate variability in a collegiate female soccer team. J Strength Cond Res. 2016;30(2):378-85. doi:10.1519/jsc.0000000000001095.

110. Gabrys T, Stec K, Michalski C, Pilis W, Pilis K, Witkowski Z. Diagnostic value of Beep and Yo-Yo tests in assessing physical performance of female soccer players. Biomed Hum Kinet. 2019;11(1):110-4.

111. Hasegawa N, Kuzuhura K. Physical characteristics of collegiate women’s football players. Football Science. 2015;12:51-7.

112. Krustrup P, Mohr M, Ellingsgaard H, Bangsbo J. Physical demands during an elite female soccer game: importance of training status. Med Sci Sports Exerc. 2005;37(7):1242-8. doi:10.1249/01.mss.0000170062.73981.94.

113. Martinez-Lagunas V, Hartmann U. Validity of the Yo-Yo Intermittent Recovery Test Level 1 for direct measurement or indirect estimation of maximal oxygen uptake in female soccer players. Int J Sports Physiol Perform. 2014;9(5):825-31. doi:10.1123/ijspp.2013-0313.

114. Morales J, Roman V, Yanez A, Solana-Tramunt M, Alamo J, Figuls A. Physiological and psychological changes at the end of the soccer season in elite female athletes. Journal of human kinetics. 2019;66:99-109. doi:10.2478/hukin-2018-0051.

115. Schmitz RJ, Cone JC, Tritsch AJ, Pye ML, Montgomery MM, Henson RA et al. Changes in drop-jump landing biomechanics during prolonged intermittent exercise. Sports health. 2014;6(2):128-35. doi:10.1177/1941738113503286.

116. Scott D, Lovell R. Individualisation of speed thresholds does not enhance the dose-response determination in football training. J Sports Sci. 2018;36(13):1523-32. doi:10.1080/02640414.2017.1398894.

117. Sjokvist J, Laurent MC, Richardson M, Curtner-Smith M, Holmberg HC, Bishop PA. Recovery from high-intensity training sessions in female soccer players. J Strength Cond Res. 2011;25(6):1726-35. doi:10.1519/JSC.0b013e3181e06de8.
118. Tounsi M, Jaafar H, Aloui A, Souissi N. Soccer-related performance in eumenorrheic Tunisian high-level soccer players: effects of menstrual cycle phase and moment of day. J Sports Med Phys Fitness. 2018;58(4):497-502. doi:10.23736/s0022-4707.17.06958-4.
119. Wright MD, Hurst C, Taylor JM. Contrasting effects of a mixed-methods high-intensity interval training intervention in girl football players. J Sports Sci. 2016;34(19):1808-15. doi:10.1080/02640414.2016.1139163.
120. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol. 2001;28(4):892-903.
121. Gruijters SLK, Peters GJY. Meaningful change definitions: sample size planning for experimental intervention research. Psychol Health. 2020:1-16. doi:10.1080/08870446.2020.1841762.
122. Lakens D. Sample Size Justification. 2021. https://doi.org/10.31234/osf.io/9d3yf.
123. Cook JA, Hislop J, Adewuyi TE, Harrild K, Altman DG, Ramsay CR et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. Health Technol Assess. 2014;18(28):v-vi, 1-175. doi:10.3310/hta18280.
124. Terwee CB, Terluin B, Knol DL, de Vet HC. Combining clinical relevance and statistical significance for evaluating quality of life changes in the individual patient. J Clin Epidemiol. 2011;64(12):1465-7; author reply 7-8. doi:10.1016/j.jclinepi.2011.06.015.
125. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates. p. 567; 1988.
126. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124.
127. Watt JA, Veroniki AA, Tricco AC, Straus SE. Using a distribution-based approach and systematic review methods to derive minimum clinically important differences. BMC Med Res Methodol. 2021;21(1):41. doi:10.1186/s12874-021-01228-7.
128. Cella D, Bullinger M, Scott C, Barofsky I. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc. 2002;77(4):384-92. doi:10.4065/77.4.384.
129. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 2010;63(1):37-45. doi:10.1016/j.jclinepi.2009.03.011.
130. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol. 2001;54(12):1204-17. doi:10.1016/s0895-4356(01)00407-3.
131. Redelmeier DA, Tversky A. Discrepancy between medical decisions for individual patients and for groups. N Engl J Med. 1990;322(16):1162-4. doi:10.1056/nejm199004193221620.
132. Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. J Pain Symptom Manage. 2002;24(6):547-61. doi:10.1016/s0885-3924(02)00529-8.
133. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ. 2020;369:m1714. doi:10.1136/bmj.m1714.

134. Impellizzeri FM, Rampinini E, Maffiuletti NA, Castagna C, Bizzini M, Wisloff U. Effects of aerobic training on the exercise-induced decline in short-passing ability in junior soccer players. Appl Physiol Nutr Metab. 2008;33(6):1192-8. doi:10.1139/H08-111.
135. Draak THP, de Greef BTA, Faber CG, Merkies ISJ. The minimum clinically important difference: which direction to take. Eur J Neurol. 2019;26(6):850-5. doi:10.1111/ene.13941.
136. Higgins JPT, Green S, (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from https://handbook-5-1.cochrane.org/chapter_16/16_5_4_how_to_include_multiple_groups_from_one_study.htm. 2011.
137. Hislop J, Adewuyi TE, Vale LD, Harrild K, Fraser C, Gurung T et al. Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review. PLoS Med. 2014;11(5):e1001645. doi:10.1371/journal.pmed.1001645.
138. Tenan MS, Simon JE, Robins RJ, Lee I, Sheean A, Dickens JF. Anchored minimal clinically important difference metrics are biased by regression-to-the-mean. J Athl Train. 2020. doi:10.4085/1062-6050-0368.20.
139. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. J Clin Epidemiol. 2017;89:188-98. doi:10.1016/j.jclinepi.2017.06.009.

List of Figures and Tables

Figure 1. Flow diagram of the systematic review process for linear speed (5-m and 30-m)
Figure 2. Flow diagram of the systematic review process for CMJ
Figure 3. Flow diagram of the systematic review process for Yo-Yo IR1
Figure 4. Raincloud plots for data distribution and degree of measurement error from the test-retest data (a) 5-m sprinting, (b) 30-m sprinting, (c) CMJ, and (d) Yo-Yo IR1
Figure 5. Plots illustrating the mean (95%CI) for the results of change values deemed of practical relevance by practitioners (survey data), the minimal detectable change (test-retest analysis) and the evidence synthesis (τ) for (a) 5-m sprinting, (b) 30-m sprinting, (c) CMJ, and (d) Yo-Yo IR1.

Table 1. Study eligibility criteria

Supplementary Table 1. Database search strategy
Supplementary Table 2. Relative quality of meta-analytical models for 5-m sprinting time data
Supplementary Table 3. Relative quality of meta-analytical models for 30-m sprinting time data
Supplementary Table 4. Relative quality of meta-analytical models for CMJ height data
Supplementary Table 5. Relative quality of meta-analytical models for Yo-Yo IR1 distance data
Supplementary Table 6. Practically relevant changes in physical performance measures survey questions
Supplemetary Table 7. Analysis code