A Bayesian approach to interpret intervention effectiveness in strength and conditioning: Part 2. Effect size selection and application of Bayesian updating.

Review Article

Running head: Bayesian updating of S&C interventions

Paul Alan Swinton[1], Katherine Burgess[1], Andy Hall[1], Leon Greig[1], John Psyllas[1], Rodrigo Aspe[1], Patrick Maughan[1,2], Andrew Murphy[1]

Doi: 10.51224/SRXIV.11

SportRχiv hosted preprint version 1

09/09/2021

PREPRINT - NOT PEER REVIEWED

Institutions

[1] School of Health Sciences, Robert Gordon University, Aberdeen, UK

[2] College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

Corresponding Author

Dr. Paul Swinton

School of Health Sciences, Robert Gordon University

Garthdee Road

Aberdeen, UK,

AB10 7QG

p.swinton@rgu.ac.uk, +44 (0) 1224 262 3361

Please cite this paper as: Swinton, PA. Burges, K. Hall, A. Greig L. Psyllas J. Aspe R. Maughan P. Murphy A. A Bayesian approach to interpret intervention effectiveness in strength and conditioning: Part 2. Effect size selection and application of Bayesian updating. Pre-print available from SportRχiv. https://doi.org/10.51224/SRXIV.11.

Abstract

Background Effect sizes are commonly used to assess the effectiveness of interventions in strength and conditioning (S&C). The purposes of this large meta-analysis were to investigate the properties of two different effect size statistics and synthesize the large amount of data available in the form of informative Bayesian priors to quantify effectiveness of future S&C interventions.

Methods An online database and hand search of published and unpublished S&C intervention studies from the 1950's onwards was conducted. Pre- and post-intervention data comprising means and standard deviations were extracted from outcomes categorized as: maximum strength, jump performance or sprint performance. Standardised mean difference (SMD_{pre}) and percentage improvement (%Improve) obtained from the response ratio were calculated and modelled with 4-level Bayesian hierarchical meta-analysis models. Results were also used to create normally distributed priors which were incorporated into an accessible tool for assessing the effectiveness of future S&C interventions through the use of Bayesian updating.

Results Data from 628 studies comprising 5468 effect sizes were included in the analyses. Large differences were identified in the effect size distributions for maximum strength (pooled means: SMD_{pre} =0.68 [95%CrI: 0.63 to 0.73]; %Improve = 14.3% [95%CrI: 13.3 to 15.4]) and sprint performance (pooled means: SMD_{pre} =0.46 [95%CrI: 0.43 to 0.50]; %Improve = 6.8% [95%CrI: 6.3 to 7.3]). These differences were also reflected in development of Bayesian priors with the lowest means and largest relative variance obtained for sprint performance reflecting lower improvements in general, but also greater relative dispersion of results. Analysis of the tails of the effect size distributions indicated consistent overestimations of SMD_{pre} values, likely caused by underestimated standard deviations.

Conclusions Future evaluations of S&C interventions are likely to be better performed and contextualised using Bayesian approaches featuring the information and informative priors developed in this meta-analysis. To facilitate an uptake of Bayesian methods within S&C, an easily

accessible tool employing intuitive Bayesian updating was created. It is recommended that researchers and practitioners use the tool alongside the S&C specific threshold values, instead of continual isolated effect size calculations and Cohen's generic values when evaluating the effectiveness of future S&C interventions. Researchers may choose to evaluate interventions using both SMD_{pre} and percent improvement statistics given their different strengths and limitations.

Key Words: S&C; evaluation; effect size; Bayesian; prior; percent change; percent improvement

1.0 Introduction

Evidence synthesis approaches including the use of meta-analysis have proliferated in parallel with the increased volume of strength and conditioning (S&C) research. Meta-analyses in S&C frequently seek to combine relatively homogeneous studies employing similar designs, each of which address a specific research question. Examples include meta-analyses investigating repetition duration (1), weekly set volume (2,3) or periodized versus non-periodized training (4) to increase physical qualities such as strength, power or muscular hypertrophy. Where less homogenous studies are included, meta-analyses frequently incorporate more sophisticated models to account for potential confounding of study-level moderators (3). However, criticisms of meta-analyses and their ability to synthesise data from heterogenous studies have been made, with concerns primarily related to their ability to generate relevant practical applications (5). In contrast to narrow evidence synthesis approaches, large-scale meta-analyses have also been conducted in S&C to describe general trends and identify the most influential factors determining intervention effectiveness. Seminal work by Rhea and colleagues (6,7) focussed on strength training and demonstrated large differences in the expected response between untrained and highly trained participants. More recently, Swinton et al (8) showed large differences in the magnitude of change across an intervention depending on the outcome type (e.g. strength, speed, power), the intervention type (e.g. resistance, plyometric, sprint), intervention duration, training status, gender, and the degree of specificity between training and outcomes.

Of central importance to both the rigour and interpretation of meta-analyses is the choice of effect size. Most previous meta-analyses conducted in S&C have used the pre-standardised mean difference (SMD_{pre}), dividing the mean change by the pre-intervention standard deviation. One of the primary reasons for the widespread use of the SMD_{pre} includes the existence of common threshold values to apply qualitative labels describing intervention effectiveness as “small”,

"medium", or "large". However, threshold values have generally used Cohen's initial suggestions (9) which were determined arbitrarily with the behavioural and social sciences in mind, thereby presenting a limitation. To address this, Swinton et al (8) used contemporary meta-analysis methods to develop S&C specific SMD_{pre} threshold values with suggested ranges to account for moderating factors such as intervention duration, training status, gender, and the degree of specificity between training and outcome. Additional reasons for the widespread use of the SMD_{pre} includes several conceptual advantages (10) enabling description of how future individuals performing the intervention should be expected to change relative to the population form which the sample was drawn. Assuming the outcome of interest follows a normal distribution in the population, an intervention with $SMD_{pre} = 0.5$ indicates an average improvement of a half standard deviation. Therefore, an individual starting at the 25^{th} or 50^{th} percentile, should be expected to move to the 43^{rd} or 69^{th} percentile, respectively. Whilst some researchers have argued that this perspective is not the most relevant when considering response to training interventions (11), it has also been argued that the most important limitations of the SMD_{pre} reflects concerns related to reliability and sampling (12). Most research conducted in S&C is nomothetic where interest lies beyond the specific sample investigated and the aim is to generalise findings to the larger population (13). To obtain sample statistics that provide valid statistical inference, the SMD_{pre} requires both the mean change and pre-intervention standard deviation to be representative of the population. This can be achieved on average through random sampling of participants from the target population (10). However, random sampling is rarely performed in S&C, and instead convenience samples are routinely obtained from a single sports team or small selection of teams. In contrast to the homogeneity of many samples, it can be argued that populations of interest within S&C are generally heterogeneous, with focus predominantly on the training status (e.g. untrained, recreationally trained and highly trained) and type of sporting activity performed (e.g. strength sports, field sports, collision sports, endurance sports). Given diffuse populations and frequent use of convenience samples representing restrictive sub-sections, it should be expected

that pre-intervention sample standard deviations will in general underestimate this population parameter, and thereby lead to an overestimation of the population effect size. Where very restrictive and overly homogenous samples are recruited, extremely large and physiologically implausible SMD_{pre} may be obtained. Findings supporting this position were reported by Swinton et al (8) where over 100 outliers were identified, including many SMD_{pre} values greater than 10. Whilst the use of convenience samples may also lead to biased means that either over- or underestimate population values, this may not present as large a limitation. If there is no or a weak relationship between pre intervention values and the change score, the sample mean difference may still provide an appropriate estimate of the mean population change. Additionally, across the entire research base it may be expected that studies will include non-random samples with both negatively and positively biased means. In contrast, for SMD_{pre} we should expect convenience and thereby frequent restricted samples to overestimate the population value.

An alternative effect size focused on sample means that can provide simple and intuitive interpretations of the magnitude of an intervention effect is the ratio of means (14). Like the SMD_{pre}, the ratio of means (post-intervention mean divided by pre-intervention mean) is dimensionless enabling synthesis of outcomes across different units and scales. It has been argued that the ratio of means, which can also be interpreted in terms of percentage improvement (e.g. 1.50 is equivalent to a 50% increase from baseline and 0.8 is equivalent to a 20% decrease) is easier to interpret than the SMD_{pre} making it a more applicable summary statistic (15). When working with the ratio of means, the natural logarithm is generally used for statistical analyses before back transforming to interpret results. The natural logarithm of the ratio of means is commonly referred to as the response ratio (RR) and in some disciplines such as ecology is the most popular effect size metric for both individual studies and meta-analyses (16). The RR like the SMD_{pre} has received substantive statistical investigation with adjustments identified to account for issues such as small-

sample bias and accurate sampling variance to improve properties for inclusion in meta-analyses (17,18). The RR has received limited use in previous meta-analyses conducted in sport science (19). However, it has been recommended that interventions to improve physical performance should be meta-analysed in percent units to better communicate results (20,21). As a result, several previous meta-analyses have converted mean difference effect sizes to percentage change and obtained standard errors from inferential statistics to conduct meta-analytic models (21-23). Additionally, a previous meta-analysis investigating the dose-response relationship between training volume and increases in muscle mass (3) performed a traditional analysis with SMD_{pre} but reported equivalent percentage gain values to better communicate results. Given the clear interest in interpreting results in a scale other than standard deviation units and the limitations that are likely to exist with estimates of population standard deviations, percentage improvement calculated from the RR with its established statistical properties represents an appropriate alternative for meta-analyses in S&C.

Large meta-analyses also have the potential to assist researchers and practitioners quantifying the effectiveness of future interventions. Almost all statistical analyses conducted in S&C research employ a frequentist framework where effect sizes are calculated anew without including prior information regarding likely values based on previous research. In the minority of cases, where uncertainty in effect sizes are quantified, confidence intervals are used which frequently suffer from misinterpretation and non-intuitive interpretations (24). Additionally, due to the small sample-sizes generally included in S&C interventions (8), uncertainty in effect sizes calculated under a frequentist framework are likely to lack precision. Instead, a Bayesian framework enables individuals to include prior information and express the uncertainty of effect sizes in an intuitive probabilistic manner (e.g. using a posterior distribution), borrowing strength from previous research to increase precision. Common critiques of Bayesian approaches include the complexity

that may exist with the analysis process and challenges in creating suitable informative priors (25). However, meta-analysis models estimate a set of intervention effects rather than a single estimate and can thus be used to develop priors which combine with new data using simple calculations to obtain a normally distributed posterior describing the most likely population effect size (26). Therefore, the purpose of this study was to build upon the meta-analysis of Swinton et al (8) and compare the SMD_{pre} and percentage improvement effect sizes to quantify intervention effectiveness in S&C. A focus of the analysis was placed on the influence of the standard deviation in determining SMD_{pre} values. In addition, meta-analyses were used to generate informative priors and threshold values to calculate Bayesian posteriors and effectively interpret future S&C interventions.

2.0 Methods

2.1 Search strategy

Studies obtained were part of a search for a previous meta-analysis (8) comprising published and unpublished research in the English language that included S&C interventions conducted prior to January 2018. The search was performed using Embase, Medline, Web of Science, Sport Discus and Google Scholar. Hand searching of relevant journals including Medicine and Science in Sports and Exercise, the Journal of Strength and Conditioning Research, and Research Quarterly was also conducted. Database search terms were included to identify various training modes, longitudinal interventions, and a range of outcome measures relevant to S&C.

2.2 Inclusion criteria

Inclusion criteria comprised: 1) any S&C intervention-based study ≥ 4 weeks; 2) healthy trained or untrained participants with a mean age between 14 and 60; 3) intervention group with a minimum of 4 participants; 4) pre and post intervention means and standard deviations collected from an outcome measure identified as maximum strength, vertical jump or sprint performance. Studies comprising interventions that were predominantly aerobic-based or rehabilitation focused were excluded.

2.3 Study selection and data extraction

Following deduplication, a three-level selection process comprising title, then abstract then full-text screening was completed. Studies were screened and selected for inclusion independently by AM with discussions with PS and KB where required. A standardised extraction codebook was developed using Microsoft Excel, with data extracted and coded independently by four researchers (AM, JP, AH, LG) in duplicate with AM completing extraction for all studies to provide

consistency. Maximum strength outcomes included a measure of maximum force production where time was not limited (e.g. 1-6 repetition maximum, isometric mid-thigh pull, peak torque). Jump performance outcomes included jump tests where a measure of vertical jump height or distance was collected. Sprint performance outcomes included measurement of the time to complete a specified linear distance or the velocity achieved. To investigate the variation in baseline standard deviation across studies, a sub-selection of the most popular tests in each outcome category were identified enabling these outcomes to be analysed in the same absolute scale. These included 1RM tests in the squat and bench press (measured in kg), vertical squat and countermovement jumps (measured in cm), and time to sprint 10 m, 20 m, 30 m, 40 m and 40 yds (measured in seconds). Training status was categorized using definitions previously set by Rhea (30) based on S&C training experience: untrained (<1 year); recreationally trained (1-5 years); highly trained (>5 years). Where pre-post intervention data were not presented in text but in figures, data were extracted using PlotDigitizer 2.6.8 Windows.

2.4 Statistical analysis

Effect sizes and their sampling variance were calculated using group mean and standard deviation values calculated pre-intervention and at any subsequent time-point. The SMD_{pre} and RR effect sizes and their within-study variances σ_e^2 were calculated using the following formulae:

$$\mathrm{SMD_{pre}} = \left(1 - \frac{3}{4n-5}\right)\left(\frac{\bar{x}_{Post} - \bar{x}_{Pre}}{Sd_{Pre}}\right)$$

where n is the number of participants in the intervention and the first term comprises a small-study bias term $c(n-1)$.

$$\sigma_e^2\left(\mathrm{SMD_{pre}}\right) = (c(n-1)^2)\left(\frac{n-1}{n(n-3)}\right)\left(2(1-r) + n\mathrm{SMD_{pre}^2}\right) - \mathrm{SMD_{pre}^2}$$

where r is the correlation between repeated measures.

$$RR = \ln\left(\frac{\bar{x}_{Post}}{\bar{x}_{Pre}}\right) + \frac{1}{2}\left(\frac{Sd_{Post}^2}{n\bar{x}_{Post}^2} - \frac{Sd_{Pre}^2}{n\bar{x}_{Pre}^2}\right)$$

$$\sigma_e^2(RR) = \frac{Sd_{Post}^2}{n\bar{x}_{Post}^2} + \frac{Sd_{Pre}^2}{n\bar{x}_{Pre}^2} - \frac{2rSd_{Post}Sd_{Pre}}{n\bar{x}_{Post}\bar{x}_{Pre}}$$

Percentage improvement (e.g. positive value represents improvement and negative value represents a decline in performance) was then calculated using the following formulae depending on whether an increase or decrease in the outcome represented an improvement in performance.

$$\%\text{Improve} =_{+}^{-} \{100(\exp(RR) - 1), 100(1 - \exp(RR))\}$$

Relationships between baseline mean and baseline standard deviation were investigated using log-log linear regression for outcomes measured on different scales, and standard linear regression with sub-analyses of the most common tests measured on the same scale. Prior to conducting full meta-analysis models, the tails of the empirical distributions were investigated by focusing on the smallest 1, 2 and 5% (0.005-, 0.0125-, and 0.025-quantile) and largest 1, 2 and 5% (0.975-, 0.9875, and 0.995-quantile) values for both the SMD_{pre} and percentage improvement effect sizes. Additionally, the ratio of the baseline mean and the baseline standard deviation was also calculated at the tails of the empirical distribution for the sub-selected most common tests.

All meta-analyses were conducted using a nested four-level Bayesian mixed effects meta-analytic model (8). The series of nestings included the individual study (level 4), the outcome (level 3), the measurement occasion (level 2) and the sampling variance (level 1). To account for uncertainty in σ_e^2 due to non-reporting of r, the values were allowed to vary and were estimated by including an informative Gaussian prior approximating correlation values centred on 0.7 and ranging from 0.5 to 0.9. Variance partition coefficients (VPCs) were used to quantify the relative variance explained

across the different levels of the hierarchy, with addition of VPCs used to estimate the expected (population) correlation between two randomly chosen elements within the same nesting structure (27). The parameters obtained from the meta-analysis models were then used to calculate small, medium and large threshold values for each of the outcome types. This was achieved by generating posterior predictions from each meta-analysis model and calculating the 0.25-, 0.5-, and 0.75-quantiles (8). Weakly informative Student-t prior and half-t priors with 3 degrees of freedom and scale parameter equal to 2.5 were used for intercept and variance parameters (28). Outlier values were identified by adjusting the empirical distribution by a Tukey g-and-h distribution and obtaining the 0.0035- and 0.9965-quantiles, with values beyond these points removed prior to further analysis (29). Meta-analyses were performed using the R wrapper package brms interfaced with Stan to perform sampling (30). Convergence of parameter estimates were obtained for all models with Gelman-Rubin R-hat values below 1.1 (31).

To build prior distributions for each outcome type, the posterior mean and standard deviation (calculated as the square root of the sum of variance components across levels 2 to 4) obtained from the meta-analysis models along with their credible intervals (mean: 0.025 to 0.975-quantile; standard deviation: 0.125 to 0.875-quantile) were collected. An expanded grid optimisation search was then used to select a mean and standard deviation value to represent the normally distributed prior $\left(\theta \sim \text{Normal}(\theta_0, \sigma_0^2)\right)$ across the credible intervals identified. For each point on the grid, the mean and standard deviation value was used to calculate the quantile value of the small, medium and large thresholds previously identified. A least squares approach was then used with the cost function equal to the squared sum of the differences between the quantile values collected and the corresponding 0.25, 0.5 and 0.75 reference values. Finally, a supplementary file was created so that the prior distributions calculated could be combined with data from future S&C interventions to produce posterior distributions and probabilistic information on whether the intervention exceeds

the context specific small, medium and large thresholds. With new data, the Bayesian updating is achieved by calculating the effect size ES_{new} and standard error $\sigma_{e_{new}}$. The standard error is then transformed into a standard deviation of the participant level outcome σ using

$$\sigma^2_{\text{e}_{\text{new}}} = \frac{\sigma^2}{n_{new}}$$

where n_{new} is the number of participants in the intervention of interest. The prior variance σ_0^2 is then re-expressed so that the amount of information contained in the prior distribution is equivalent to an intervention with n_0 participants where

$$\sigma_0^2 = \frac{\sigma^2}{n_0}.$$

The Bayesian updating for the posterior distribution of the effect size θ is then achieved by using the following formula (26)

$$\theta|ES_{new} \sim \text{Normal}\left(\frac{n_0\theta_0 + n_{new}ES_{new}}{n_0 + n_{new}}, \frac{\sigma^2}{n_0 + n_{new}}\right).$$

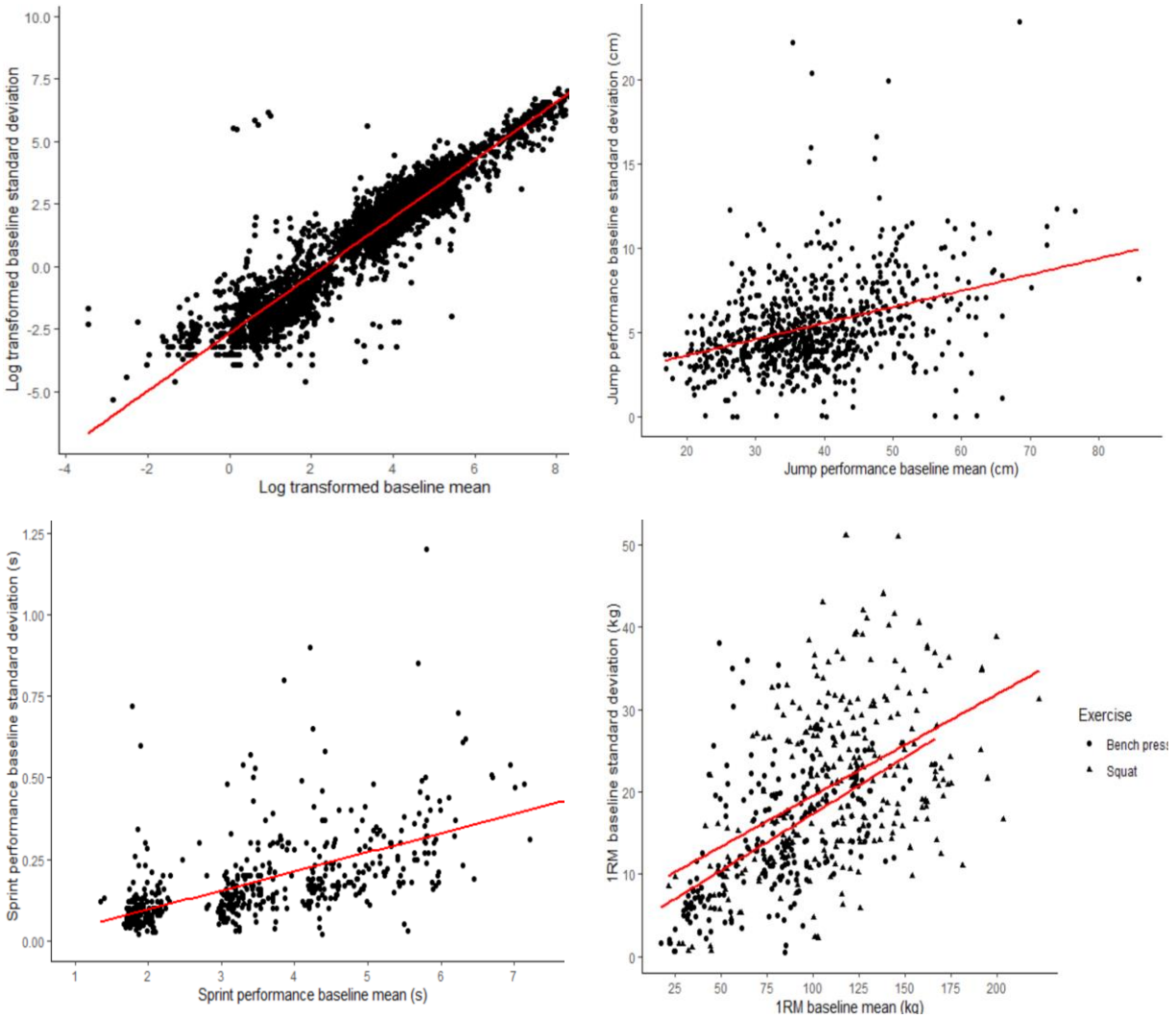
3.0 Results

The search strategy employed for the original meta-analysis (Swinton et al. 2021) returned 110,662 records which reduced to 2108 studies following deduplication and title screening. This reduced to 973, and 706 studies following abstract and full-text screening, respectively. A total of 628 studies featured the required data to be included in the present meta-analysis with most studies comprised untrained participants (n=374, 59.6%), followed by recreationally trained (n=212, 33.8%) then highly trained (n=42, 6.7%). A total of 2632 maximum strength effect sizes were extracted from 421 studies, followed by 1574 jump performance effect sizes from 382 studies, and 1262 sprint performance effect sizes from 257 studies. When restricting outcomes to sub-analyses of the most common tests from each category, a total of 957 vertical jump performance (unloaded squat and counter-movement) effect sizes were extracted from 339 studies, followed by 607 sprint performance (10 m, 20 m, 30 m, 40 m and 40 yrd sprint times) effect sizes from 180 studies, 344 1RM squat effect sizes from 135 studies and 318 1RM bench press effect sizes from 110 studies.

Analyses of the relationships between baseline standard deviation and baseline mean showed functional differences in form across tests measured on different scales (e.g. vertical jump cm and peak force N), and differences in relative magnitudes (e.g. ratio values) across different outcome domains. Power relationships were identified when including tests across different scales as identified by a linear log-log plot for the complete data set (Figure 1A) and within each domain (not shown). In contrast, standard linear relationships appeared suitable for the sub-analyses of jump performance measured in centimetres, sprint times measured in seconds, and 1RM squat and bench press tests measured in kilograms (Figures 1B-1D). Simple linear regression conducted on the sub-analyses of the most common tests identified large variations in standard deviations when regressed on the baseline mean, with most values (e.g. 4 times the standard error) falling within a

range of 10.3 cm for vertical jump performance, 0.45 s for sprint performance, and 24.1 and 34.8 kg for the 1RM bench press and squat, respectively.

Figure 1: Relationships between baseline means and standard deviation. 1A (top-left): Log-log transformations across the whole data set. 1B (top-right): Jump performance measured in centimetres. 1C (bottom-left): Sprint performance measured in seconds. 1D (bottom-right): 1RM squat and bench press performance measured in kg.



Standard error values for linear regression performed on jump, sprint, bench press and squat data were: 2.6 cm, 0.11 s, 6.0 kg and 8.7 kg, respectively.

Direct calculation of the mean and standard deviation of effect size statistics from the complete empirical data returned SMD_{pre} values of 0.79±1.6 and percentage improvement values of 10.9±14.4%. The same calculations applied to each of the outcome domains returned SMD_{pre} values of 0.95±1.7, 0.62±1.6 and 0.63±1.6; and percentage improvement values of 16.0±17.3%, 7.6±7.8% and 2.4±3.4% for maximum strength, jump and sprint performance, respectively. Prior to applying the meta-analytic model, the tails of the empirical values were investigated (Table 1) and demonstrated long right tails with relatively similar SMD_{pre} values obtained across the different domains (i.e., similar large physiologically implausible values). In contrast, substantive differences were obtained for percentage improvement with more extreme large values obtained for maximum strength, followed by jump then sprint performance. As a final check of the tails of the distribution, the ratios of the baseline standard deviation relative to the mean were calculated for sub-analyses of the most common tests (Table 2). Similar patterns were obtained across all tests with relatively small changes in the ratio for the left tail as compared with ratios from values at the centre. In contrast, large changes were obtained in the right tail with substantively and progressively lower ratios progressing further into the tail. The largest changes were obtained for the squat where the baseline standard deviation was estimated to equal approximately 23% of the baseline mean for effect sizes close to the median, which reduced to 14, 8 and then 3% for effect sizes close to the 0.975-, 0.9875- and 0.995-quantiles, respectively.

Table 1: Direct calculation of largest and smallest 1, 2 and 5% of effect sizes across outcome types.

		0.005-Quantile	**0.0125-Quantile**	**0.025-Quantile**	**0.5-Quantile**	**0.975-Quantile**	**0.9875-Quantile**	**0.995-Quantile**
Outcome	**Statistic**							
All	SMD_{pre}	-0.84	-0.50	-0.29	0.52	3.3	5.3	9.1
	%Improve	-14.3%	-6.6%	-3.6%	7.1%	46.8%	61.3%	93.4%
Maximum Strength	SMD_{pre}	-0.84	-0.42	-0.23	0.62	4.0	5.9	8.6
	%Improve	-17.0%	-9.1%	-4.9%	12.5%	59.7%	84.0%	113%
Jump performance	SMD_{pre}	-0.74	-0.38	-0.25	0.47	2.5	3.3	7.8
	%Improve	-11.1%	-5.8%	-3.4%	6.3%	27.6%	34.7%	42.1%
Sprint Performance	SMD_{pre}	-0.88	-0.68	-0.44	0.36	2.6	5.5	11.7
	%Improve	-6.9%	-3.9%	-2.6%	2.0%	10.8%	12.4%	16.0%

Table 2: Direct calculation of the percentage of the baseline standard deviation relative to the mean in the tails of empirical distributions (largest and smallest 1, 2 and 5% of effect sizes).

	0.005-Quantile	**0.0125-Quantile**	**0.025-Quantile**	**0.5-Quantile**	**0.975-Quantile**	**0.9875-Quantile**	**0.995-Quantile**
Outcome							
Jump Performance	9.5%	12.1%	14.1%	15.3%	9.5%	5.6%	0.4%
Sprint Performance	2.7%	3.9%	4.4%	5.6%	2.8%	0.9%	0.5%
1RM Squat	17.9%	21.1%	23.0%	23.4%	14.0%	7.8%	2.5%
1RM Bench press	14.8%	15.3%	16.1%	16.1%	9.8%	6.5%	3.8%

Prior to meta-analysis, a total of 106 outliers were removed with lower bound SMD_{pre} and percentage improvement thresholds of -0.91 and -11.4%, and upper bound SMD_{pre} and percentage improvement thresholds of 6.8 and 93.4%. Based on shrinkage from application of the meta-analysis model, and borrowing of information across studies and outcomes, the pooled mean estimate obtained across all domains was reduced to $SMD_{pre0.5}$=0.56 [95%CrI: 0.53 to 0.59] and $\%Improve_{0.5}$=9.3 [95%CrI: 8.7 to 9.9%]. Estimates of the total standard deviation (calculated from summation of levels 2, 3 and 4) were $\sigma(SMD_{pre})_{0.5}$=0.42 [85%CrI: 0.40 to 0.45] and $\sigma(\%Improve)_{0.5}$=8.5% [85%CrI: 8.0 to 09.2]. Effect size statistics; small, medium and large thresholds; and variance parameters for maximum strength, jump and sprint performance are presented in table 3. The results showed large differences across outcome types with the greatest effect sizes obtained for maximum strength and substantively smaller effect sizes obtained for sprint performance. Mean and standard deviation values for future prior distributions are presented in table 4. The supplementary file includes prior distributions for each of the domains and includes calculations presented in the statistical analysis section to generate posterior distribution parameters and probability of exceeding small, medium and large thresholds when a user enters either raw data (individual pre- and post-intervention) or summary data (sample size, pre- and post-intervention mean and standard deviation).

Table 3: Meta-analysis results for all pooled outcomes and domain specific outcomes.

		Mean [95% CrI]	Small (0.25-quantile) [95% CrI]	Medium (0.5-quantile) [95% CrI]	Large (0.75-quantile) [95% CrI]	Study Level VPC [75% CrI]	Outcome Level VPC [75% CrI]	Measurement Occasion VPC [75% CrI]
Outcome	**Statistic**							
All	SMD_{pre}	0.56 [0.53 to 0.59]	0.18 [0.17 to 0.20]	0.49 [0.47 to 0.51]	0.84 [0.82 to 0.86]	0.27 [0.26 to 0.28]	0.14 [0.11 to 0.17]	0.01 [0.00 to 0.03]
	%Improve	9.3 [8.7 to 9.9]	2.6 [2.3 to 2.8]	8.1 [7.7 to 8.4]	15.6 [15.3 to 16.1]	0.30 [0.29 to 0.31]	0.29 [0.24 to 0.34]	0.01 [0.00 to 0.06]
Maximum Strength	SMD_{pre}	0.68 [0.63 to 0.73]	0.25 [0.22 to 0.27]	0.60 [0.57 to 0.62]	0.99 [0.96 to 1.0]	0.28 [0.27 to 0.29]	0.10 [0.07 to 0.13]	0.02 [0.00 to 0.04]
	%Improve	14.3 [13.3 to 15.4]	6.0 [5.4 to 6.5]	13.6 [13.0 to 14.3]	22.7 [22.0 to 23.4]	0.30 [0.29 to 0.30]	0.21 [0.15 to 0.26]	0.06 [0.00 to 0.16]
Jump Performance	SMD_{pre}	0.46 [0.43 to 0.50]	0.18 [0.15 to 0.20]	0.44 [0.42 to 0.47]	0.73 [0.70 to 0.76]	0.26 [0.25 to 0.28]	0.01 [0.00 to 0.03]	0.00 [0.00 to 0.02]
	%Improve	6.8 [6.3 to 7.3]	2.8 [2.4 to 3.2]	6.7 [6.3 to 7.1]	11.0 [10.6 to 11.5]	0.28 [0.27 to 0.29]	0.06 [0.03 to 0.09]	0.03 [0.00 to 0.11]
Sprint Performance	SMD_{pre}	0.41 [0.36 to 0.46]	0.08 [0.05 to 0.12]	0.36 [0.33 to 0.39]	0.66 [0.62 to 0.71]	0.28 [0.27 to 0.29]	0.00 [0.00 to 0.02]	0.01 [0.00 to 0.02]
	%Improve	2.5 [2.2 to 2.8]	0.5 [0.3 to 0.7]	2.1 [2.0 to 2.3]	3.9 [3.7 to 4.2]	0.31 [0.30 to 0.31]	0.06 [0.03 to 0.09]	0.04 [0.00 to 0.09]

CrI: Credible interval. VPC: Variance partition coefficients. SMD_{pre}: Standardised mean difference using the baseline standard deviation.

Table 4: Mean and standard deviations of future prior distributions for SMD_{pre} and RR effect sizes statistics across outcome types.

		Prior Mean	Prior Standard distribution
Outcome	**Statistic**		
All	SMD_{pre}	0.53	0.45
	RR	0.083	0.077
Maximum Strength	SMD_{pre}	0.63	0.54
	RR	0.130	0.108
Jump performance	SMD_{pre}	0.45	0.36
	RR	0.066	0.050
Sprint Performance	SMD_{pre}	0.37	0.43
	RR	0.022	0.026

4.0 Discussion

The purpose of this meta-analysis was to compare the SMD_{pre} and percentage improvement effect sizes, and their ability to quantify effectiveness of S&C interventions. In addition, the meta-analysis sought to develop prior distributions for the effect sizes, so that Bayesian methods could be used to assess the effectiveness of future S&C interventions in an informative and intuitive manner. In general, the analyses showed similar findings for SMD_{pre} and percentage improvement, with the greatest effect sizes obtained for maximum strength outcomes, and a substantive decrease for sprint performance. Some differences were identified in the composition of the meta-analysis models including greater relative variances at the outcome level for percentage improvement compared with SMD_{pre}. For both effect sizes the positive tails of the empirical distribution exhibited extremely large values. These large values only occurred for maximum strength outcomes in percentage improvement (~60 to 110% improvement) but were consistently large and physiologically implausible for SMD_{pre} across all outcomes (~4 to 12). Extremely large SMD_{pre} values were likely influenced by underestimated standard deviation values with analyses demonstrating substantively lower standard deviations relative to baseline means in the right but not left tails. Development of the Bayesian prior distributions resulted in relatively large spreads with standard deviation values close to the mean, and for sprint performance standard deviations were greater in value, which was consistent with the finding that a substantive proportion of the distribution included effect sizes close to zero.

Whilst SMD_{pre} values can provide an informative means of interpreting the effectiveness of an S&C interpretation, they may not be readily interpretable for practitioners. In contrast, percentage improvement, which is obtained with a simple transform of the relative ratio statistic provides one of the most intuitive means of interpreting the magnitude of an effect and are consistent with how many conceptualise and discuss intervention effects (32). For example, in the present analysis the

greatest effect sizes were obtained for maximum strength outcomes with the SMD_{pre} small, medium and large thresholds equal to 0.18 [95%CrI: 0.17 to 0.20], 0.49 [95%CrI: 0.47 to 0.51] and 0.84 [95%CrI: 0.82 to 0.86], respectively. In contrast, expressed as a percentage improvement the thresholds were equal to 6.0% [95%CrI: 5.4 to 6.5], 13.6% [95%CrI: 13.0 to 14.3] and 22.7% [22.0 to 23.4] which are immediately more interpretable. However, the greatest conceptual difference between the two effect sizes is evident when comparing thresholds between maximum strength and sprint performance. For sprint performance the small, medium and large SMD_{pre} thresholds decrease to 0.08 [95%CrI: 0.05 to 0.12], 0.36 [0.33 to 0.39] and 0.66 [0.62 to 0.71]. These results show that a substantive proportion (~15 to 20%) of the effect size distribution are close to or below zero, whereas the large sprint performance threshold is between the medium and large maximum strength thresholds. In contrast, the percentage improvement thresholds for sprint performance were equal to 0.5% [95%CrI: 0.3 to 0.7], 2.1% [95%CrI: 2.0 to 2.3] and 3.9% [3.7 to 4.2], such that all thresholds were below even the small maximum strength threshold. These observations reflect differences in relationships between the means and standard deviations for each outcome and demonstrate the conceptual difference between effect sizes describing expectations of how participants will change their relative position within a population compared to the magnitude of the change relative to the starting value.

The potential for restricted sampling of a population to bias standardized effect sizes such as the SMD_{pre} was highlighted by Baguley (12). If the sample is a truncated sample (missing one or both tails), then the standard deviation is likely to be underestimated such that the SMD_{pre} will be positively biased. This scenario is most likely to occur in S&C research where random sampling is uncommon and often convenience samples are used, including recruitment from a single team where participants may be relatively homogenous given similar training experiences. In contrast, sampling only from the tails is likely to overestimate the standard deviation (12) and thereby

negatively bias the SMD_{pre}. In S&C, this situation may occur in studies recruiting both males and females where the outcome variable has a large sex stratification. The results from the present analysis highlight the likely role that recruitment practices have played in calculated standard deviations and thereby estimated SMD_{pre} values. To assess the range in values, the present analysis investigated the relationship between baseline mean and standard deviations. Clear power relationships were identified by linear log-log plots when analysing the whole data set (Figure 1A) or outcomes on different scales within a specific domain. However, to more intuitively quantify ranges in standard deviations, analyses were made regressing standard deviations on means measured in the most common tests recorded on the same absolute scales. Assuming linear relationships (Figure 1B-1D), the standard error from the regression analysis was used such that given a normal distribution, four times the standard error provided an estimate (~95% coverage) of the range. The results showed a range of 10.3 cm for vertical jump performance, 0.45 s for sprint performance, and 24.1 and 34.8 kg for the 1RM bench press and squat, respectively. To illustrate the effect that these ranges can have on SMD_{pre} values, if we consider a participant group of n=8, with a baseline mean of 100 kg, the regression analysis suggests a typical baseline standard deviation of 19.5 kg. If the mean improvement was equal to 10 kg, this would produce a SMD_{pre} value of 0.46 (which based on the results of this review would be considered a small to medium effect). However, based on actual differences in the population variance, or more likely inappropriate sampling, the results of the present analysis indicate that standard deviations of $19.5\pm(34.8/2)$ may be reported. Based on these differences, the SMD_{pre} value could decrease to 0.24 (considered a small effect size) or increase to 4.2 which is extremely large, and very unlikely to occur over a single training intervention. However, analysis of the positive tails of the distributions consistently demonstrated SMD_{pre} values greater than 4 for the top 2% of results (Table 1), and ratios of standard deviations and means were much smaller in the positive tails compared to other parts of the distribution (Table 2). Whilst large improvements are possible, especially with untrained participants (ref), the analyses presented here suggest that extremely large

SMD_{pre} values are likely to be inflated by inappropriate sampling. In addition, it is possible that values not in the tails but more central in the distribution reflect smaller population effect sizes that were inflated due to the same process.

One approach to obtain better estimates of effect sizes is to use Bayesian approaches. One of the primary challenges and biggest criticisms of Bayesian methods has been the selection of appropriate priors (25,33). Where substantive and relevant external information is present, attempts should be made to incorporate this within an informative prior (33). One of the most effective sources of information to build priors to better assess the effectiveness of future interventions includes meta-analyses such as that presented here (26). An additional challenge in the effective use and uptake of Bayesian methods is a lack of formal training and familiarity with the approaches (34). In the present study attempts have been made to address both challenges by firstly, creating priors that are based on a large volume of research covering the outcome domains generally featured in S&C research; and secondly, employing a relatively simple Bayesian updating method which can be understood intuitively and facilitated in software that is familiar with both researchers and practitioners (35). The method adopted expresses both the prior and posterior distribution of the effect size as normal distributions which are familiar and simple to assess overall suitability by examining stated probabilities. For example, based on the meta-analysis results obtained here, an SMD_{pre} prior with mean 0.68 and standard deviation of 0.54 was developed for maximum strength outcomes. This asserts that the prior probability of obtaining an SMD_{pre} value greater than 0 is p=0.896, the probability of obtaining an SMD_{pre} value between 0 and 0.5 is p=0.265, and the probability of obtaining an SMD_{pre} value greater than 1 is p=0.277. A researcher and practitioner can decide to alter the mean and standard deviation values if they believe that the probabilities investigated do not match up to their prior beliefs, but provide a useful initial anchor as they were designed to fit an extensive amount of data collected from S&C interventions.

Similarly, the updating process used to combine the prior information with data collected and generate a posterior distribution is also easily interpreted. Firstly, the method (26) updates the posterior mean as a weighted combination of the prior mean and the effect size calculated directly from the intervention. The weights are determined by the uncertainty in the estimate from the data, and where for example a small number of participants are investigated, the standard error will be large and therefore greater weight placed on the prior mean. The exact weights used are determined by matching the uncertainty in the new data and the prior, and translating the information contained in the prior to a single trial that can then be updated with the new data. To demonstrate how some of the potential issues discussed previously with regards to poor estimates of the standard deviation can be addressed, the example outlined above is continued. If we assume a correlation between the pre- and post-intervention scores of 0.7 (a requirement to calculate uncertainty in the estimate), then combining the extremely large SMD_{pre} value of 4.2 with the small sample size of n = 8, generates a standard error of 1.41. Based on a frequentist approach, a 95% confidence interval for the effect size would equal $4.2 \pm 1.96 \times 1.41$ giving a range of 1.4 to 7.0. However, given the small sample size and the large standard error, when updated in a Bayesian framework using the methods presented here and the equations in the statistical analysis section, the posterior mean and standard deviation are shrunk to 1.1 and 0.50, respectively. The effect size is still considered large but is now more plausible and can be interpreted probabilistically given the normal distribution and posterior parameters estimated (e.g. probability of at least a small effect: p=0.953; probability of at least a medium effect: p=0.840; and probability of at least a large effect: p=0.577). Note, if the sample size was much larger, say n=100, then the directly calculated effect size increases to 4.7 (due to a reduction in the bias offset) and the posterior mean is only shrunk to 3.5, as there is less uncertainty in the original estimate. This example also highlights the challenge in obtaining accurate estimates of population parameters if sampling is limited.

5.0 Conclusion

To assist practitioners in selecting and developing interventions using evidence-based practices, it is important that processes and tools are available to compare and appropriately interpret differences in results disseminated in research. Currently, the use of effect size statistics provides the most practical method of ranking interventions and determining which are most likely to provide a basis for the greatest improvements within a given population. There are, however, multiple effect size statistics that can be used, each with their own strengths and weaknesses. In S&C, the most established effect size statistic is the SMD_{pre} value which can be informative, but can provide biased results, particularly overestimations when calculated on a restricted sample of the population. An alternative effect size statistic that may fit more intuitively with practitioners' perceptions of training interventions and expectations of changes in outcome variables is percent improvement. This effect size statistic can be calculated using the response ratio which is popular in many other disciplines and whose properties are well understood (17,18). However, the response ratio has multiple limitations including its requirement to work with logarithms and challenges presented when used with outcomes that change sign (e.g. positive to negative), can equal zero, or are measured as proportions (32). Regardless of the effect size statistic used, when evaluating previous research in S&C, clear patterns emerge and large differences in distributions are evident, particularly between maximum strength and sprint performance. Knowledge that different outcome types can generate large differences in effect size distributions has several important consequences. Firstly, interpretations on the success of an intervention can be greatly influenced. For example, using previous non-S&C specific thresholds, researchers and practitioners may interpret several sprint performance interventions as being unsuitable when a more complete understanding highlights that these improvements may be relatively large and therefore the intervention appropriate to use with a given population. Secondly, knowledge of effect size distributions has important implications for setting sample size requirements for future research studies. Effect size thresholds are commonly used for power calculations using frequentist

methods and suggest that smaller sample sizes may be required for interventions aimed at developing maximum strength compared with interventions aimed at developing sprint performance. Similarly, sample size approaches using Bayesian methods can also use the prior distributions presented here in their calculations (26).

Given the large volume of S&C research and the pace at which it is accelerating, there are clear advantages to incorporating this information within future research to make better estimates, particularly where small sample sizes are common and effect sizes may be low (33). Bayesian methods are well suited to this process, and it is likely that as more disciplines and research in general take advantage of the benefits associated with Bayesian frameworks and criticisms of null hypothesis significance testing continues to grow (36), increased uptake will occur. To facilitate an increased use of Bayesian methods processes are required to address two of the main challenges which include development of appropriate priors and accessible tools and procedures that are intuitive and can be carried out ideally without need of complex software. The present study has attempted to address these challenges by developing informative priors that can be checked intuitively for their predictions. In addition, the creation of a tool in MS Excel that can perform the required calculations and generate simple, and context specific output is likely to be of benefit to both researchers and practitioners.

Acknowledgements

No funding was received for this review.

Conflicts of interest

Paul Swinton, Katherine Burgess, Andy Hall, Leon Greig, John Psyllas, Rodrigo Aspe, Patrick Maughan and Andrew Murphy declare that they have no potential conflicts of interest with the content of this article.

Author Contributions

PAS and AM designed the research. AM conducted the searches and screening. AM, LG, JP and AH extracted the data. PAS performed all statistical analyses and developed the supplementary file. PAS and AM interpreted the data analysis. PAS and AM wrote the manuscript with critical input from KB, LG, JP, AH, PM and RA. All authors read and approved the final manuscript.

References

1. References Schoenfeld BJ, Ogborn DI and Krieger JW. Effect of repetition duration during resistance training on muscle hypertrophy: A systematic review. Sports Medicine. 2015;45(4):577-585. https://doi.org/10.1007/s40279-015-0304-0
2. Ralston GW, Kilgore L, Wyatt FB and Baker JS. The effects of weekly set volume on strength gain: A meta-analysis. Sports Medicine. 2017;47(12):2585-2601. https://doi.org/10.1007/s40279-017-0762-7.
3. Schoenfeld BJ, Ogborn DI and Krieger JW. Dose-response relationship between weekly resistance training volume and increases in muscle mass: A systematic review and meta-analysis. Journal of Sports Science. 2017;35(11):11073-1082. https://doi.org/10.1080/02640414.2016.1210197.
4. Williams TD, Tolusso DV, Fedewa MV and Esco MR. Comparison of periodized and non-periodized resistance training on maximal strength: A meta-analysis. Sports Medicine. 2017; 47(10):2083-2100. https://doi.org/10.1007/s40279-017-0734-y.

5. Gentil P, Arruda A, Souza D, Giessing J, Paoli A, Fisher J and Steele J. Is there any practical application of meta-analytical results in strength training? Frontiers in Physiology. 2017. https://doi.org/10.3389/fphys.2017.00001.

6. Rhea MR, Alvar BA, Burkett LN, Ball SD. A meta-analysis to determine the dose response for strength development. Medicine & Science in Sports and Exercise. 2003;35(3):456-64. https://doi.org/10.1249/01.MSS.0000053727.63505.D4.

7. Rhea MR. Determining the magnitude of treatment effects in strength training research through the use of the effect size. Journal of Strength & Conditioning Research. 2004;18:918-20. https://doi.org/10.1519/14403.1.

8. Swinton, PA. Burges, K. Hall, A. Greig L. Psyllas J. Aspe R. Maughan P. Murphy A. A Bayesian approach to interpret intervention effectiveness in strength and conditioning: Part 1. A meta-analysis to derive context-specific thresholds. Pre-print available from SportRχiv. https://doi.org10.51224/SRXIV.9.

9. Cohen, J. Statistical power analysis for the behavioral sciences. Second Edition. Hillsdale, NJ: Lawrence Erlbaum Associate. 1988.

10. Caldwell A, Vigotsky AD. A case against default effect sizes in sport and exercise science. PeerJ 8:e10314. 2020. https://doi.org/10.7717/peerj.10314.

11. Dankel SJ and Loenneke JP. Effect sizes for paired data should use the change score variability rather than the pre-test variability. Journal of Strength and Conditioning Research. 2021;35(6):1773-1778. https://doi.org/10.1519/JSC.0000000000002946.

12. Baguley T. Standardized or simple effect size: what should be reported. British Journal of Psychology. 2009;100(3):603-617. https://doi.org/10.1348/000712608X377117.

13. Stone MH, Stone M and Sands WA. Principles and practice of resistance training. Champaign IL: Human Kinetics. 2007.

14. Hedges LV, Gurevitch J and Curtis PS. The meta-analysis of response ratios in experimental ecology. Ecology. 1999;80(4):1150-1156. https://doi.org/10.2307/177062.
15. Friedrich, JO Adhikari NKJ and Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. Journal of Clinical Epidemiology. 2011;64(5):556–564. https://doi.org/10.1016/j.jclinepi.2010.09.016.
16. Koricheva J and Gurevitch J. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology. 2014.102:828-844. https://doi.org/10.1111/1365-2745.12224.
17. Lajeunesse MJ. On the meta-analysis of response ratios for studies with correlated and multi-group designs. Ecology. 2011; 92(11):2049-2055. https://doi.org/10.2307/23034937.
18. Lajeunesse MJ. Bias and correction for the log response ratio in ecological meta-analysis. Ecology. 2015;96(8):2056-2063. https://doi.org/10.1890/14-2402.1.
19. Deb SK, Brown DR, Gough LA, Mclellan CP, Swinton PA, Sparks AS and Mcnaughton LR. Qing the effects of acute hypoxic exposure on exercise performance and capacity: A systematic review and meta-regression. European Journal of Sport Science. 2018;18(2):243-254. https://doi.org/10.1080/17461391.2017.1410233.
20. Hopkin WG, Marshall SW, Batterham AM and Hanin J. Progressive statistics for studies in sports medicine and exercise science. Medicine in Science and Sports and Exercise. 2009;41(1):3-13. https://doi.org/10.1249/MSS.0b013e31818cb278.
21. Weston M, Taylor KL, Batterham AM and Hopkins WG. Effects of low-volume high-intensity interval training (HIT) on fitness in adults: A meta-analysis of controlled and non-controlled trials. Sports Medicine. 2014;44(7):1005-1017. https://doi.org/10.1007/s40279-014-0180-z.

22. Vollard NBJ, Metcalfe RS and Williams S. Effect of number of sprints in an SIT session on change in VO2max: A meta-analysis. Medicine in Science and Sports and Exercise. 2017;49(6):1147-1156. https://doi.org/10.1249/MSS.0000000000001204.

23. Guizelini PC, de Aguiar RA, Denadai BS, Caputo F and Greco CC. Effect of resistance training on muscle strength and rate of force development in healthy older adults: A systematic review and meta-analysis. Experimental Gerontology. 2018;102:51-58. https://doi.org/10.1016/j.exger.2017.11.020.

24. Hespanhol L, Vallio CS, Costa LM and Saragiotto BT. Understanding and interpreting confidence and credible intervals around effect estimates. Brazilian Journal of Physical Therapy. 2019;23(4):290-301. https://doi.org/10.1016/j.bjpt.2018.12.006.

25. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Annals of Internal Medicine. 1999;130(12):1005-1013. https://doi.org/10.7326/0003-4819-130-12-199906150-00019.

26. Jones HE, Ades AE, Sutton AJ and Welton NJ. Use of a random effects meta-analysis in the design and analysis of a new clinical trial. Statistics in Medicine. 2018;37(30):4665-4679. https://doi.org/10.1002/sim.7948.

27. Hox, JJ, Moerbeek M, Van de Schoot R. Multilevel Analysis. Techniques and applications. 3rd edition. 2018. Routledge.

28. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006;1(3):515-34.

29. Verardi V, Vermandele C. Univariate and multivariate outlier identification for skewed or heavy-tailed distributions. The Stata Journal. 2018;18(3):517-32. https://doi.org/10.1177/1536867X1801800303.

30. Bürkner PC. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software. 2017;80(1):1-28. https://doi.org/ 10.18637/jss.v080.i01.

31. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis: Taylor & Francis; 2014.

32. Pustejovsky JE. Using response ratios for meta-analysing single-case designs with behavioral outcomes. Journal of School Psychology. 2018;68:99-112. https://doi.org/10.1016/j.jsp.2018.02.003.

33. Mengersen KL, Drovandi CC, Robert CP, Pyne DB and Gore CJ. Bayesian estimation of small effects in exercise and sports science. Plos One. 2016. https://doi.org/10.1371/journal.pone.0147311.

34. Bernards JR, Sato K, Haff GG and Bazyler CD. Current research and statistical practices in sport science and a need for a change. Sports. 2017;5(4). https://doi.org/10.3390/sports5040087.

35. Turner AN, Brazier J, Bishop C, Chavda S, Cree J, and Read P. Data analysis for strength and conditioning coaches: Using Excel to analyze reliability, differences, and relationships. Strength and Conditioning Journal. 2015;37(1):76-83. https://doi.org/10.1519/SSC.0000000000000113.

36. Wasserstein RL, Schirm AL and Lazar NA. Moving to a world beyond "p<0.05". The American Statistician. 2019. https://doi.org/10.1080/00031305.2019.1583913.