Comparative effect size distributions in strength and conditioning and implications for future research: a meta-analysis.

SWINTON, P.A. and MURPHY, A.

2024

Comparative effect size distributions in strength and conditioning and implications for future research: A meta-analysis.

Review Article

Running head: S&C comparative effect sizes

Paul Alan Swinton[1], Andrew Murphy[1,2]

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

[2] Greater Western Sydney Giants, Sydney, Australia

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

Key Words: S&C; evaluation; effect size; statistical power; sample size;

Abstract

Background Controlled experimental designs are frequently used in strength and conditioning (S&C) to determine which interventions are most effective. The purpose of this large meta-analysis was to quantify the distribution of comparative effect sizes in S&C to determine likely magnitudes and inform future research regarding sample sizes and inference methods.

Methods Baseline and follow-up data were extracted from a large database of studies comparing at least two active S&C interventions. Pairwise comparative standardised mean difference effect sizes were calculated and categorised according to the outcome domain measured. Hierarchical Bayesian meta-analyses and meta-regressions were used to model overall comparative effect size distributions and correlations, respectively. The direction of comparative effect sizes within a study were assigned arbitrarily (e.g. A vs. B, or B vs. A), with bootstrapping performed to ensure effect size distributions were symmetric and centred on zero. The middle 25, 50, and 75% of distributions were used to define small, medium, and large thresholds, respectively.

Results A total of 3874 pairwise effect sizes were obtained from 417 studies comprising 958 active interventions. Threshold values were estimated as: small = 0.14 [95%CrI: 0.12 to 0.15]; medium: = 0.29 [95%CrI: 0.28 to 0.30]; and large = 0.51 [95%CrI: 0.50 to 0.53]. No differences were identified in the threshold values across different outcome domains. Correlations ranged widely ($0.06 \leq r \leq 0.36$), but were larger when outcomes within the same outcome domain were considered.

Conclusions The finding that comparative effect sizes in S&C are typically below 0.30 and can be moderately correlated has important implications for future research. Sample sizes should be substantively increased to appropriately power controlled trials with pre-post intervention data. Alpha adjustment approaches used to control for multiple testing should account for correlations between outcomes and not assume independence.

1.0 Introduction

Strength and conditioning (S&C) is a well-established discipline within sport and exercise science that seeks to determine which interventions are most effective, and what manipulations can be made to training regimes to obtain additional improvements. In our recent large meta-analysis quantifying change following S&C interventions we found that the majority create substantive improvements across a range of outcomes (1). There is, however, less understanding of the expected differences when comparing two interventions. Interventions that have been frequently compared in S&C include variation to the training dose (e.g. single vs multiple sets (2), low volume vs high volume (3)), periodisation strategy (e.g. periodised vs non-periodised (4), linear vs undulating (5)) and training stimulus applied (e.g. low load vs high load (6), plyometric vs resistance training (7)).

The most common approach used by researchers to evaluate potential differences in S&C interventions is to perform controlled trials where the control group comprises a reference training intervention. Inferences regarding the population average treatment effect are made based on the observed differences between the sample groups. Analyses are typically conducted within a frequentist framework where uncertainty in the inference is accounted for with long-run error control by conducting null-hypothesis significance tests (NHSTs) and setting the Type 1 error rate (α) to 0.05 (8). Typically NHSTs assume that the population average treatment effect is zero (more informatively referred to as the nil-hypothesis (9)); however, the null can be set to any value including a predefined smallest worthwhile difference (10). Where NHST represents the plausibility of a specific parameter value, it is generally recommended that controlled trials also engage

in estimation that provides inferences regarding a range of plausible parameter values in the population (10). In S&C, studies frequently report the individual group standardised mean difference. This information describes how individuals are expected to move through the population after performing the specific intervention (11). These non-comparative effect sizes, however, do not describe expected differences in change between individuals performing one intervention compared to another. These potential differences are referred to as average treatment effects with controlled studies using samples to make inferences regarding the population values (12). Average treatment effects based on changes between two effective interventions such as those commonly used in S&C (1), may be more likely to be centred on, or close to zero. Additionally, comparative effect sizes provide researchers with the required information to perform a priori statistical power calculations to determine sample size for future controlled studies (13). Given the observation that relatively modest changes in the population average treatment effect can lead to substantive differences in sample size required to adequately power studies (14), it is important for S&C research to identify likely comparative effect sizes given the types of comparisons that are generally made and if this differs across outcomes measured.

A further challenge with analyses comparing S&C interventions includes the issue of multiple testing and the potential increased Type 1 error rate claiming differences in average treatment effects where none exists. Studies comparing S&C interventions tend to analyse multiple outcomes, often at multiple time points, even when outcomes selected measure the same construct (e.g. multiple outcomes each assessing maximum strength) (1). A common view is that conducting multiple statistical tests with related outcomes requires adequate lowering of the significance threshold (alpha adjustment) to control Type I errors (15). Whilst a range of alpha adjustment approaches exist, many disciplines including S&C tend to select Bonferroni corrections (16) that assume outcomes are independent (17) and can be considered overly conservative in contexts of small samples and relatively low effect sizes (18). Previous research has shown that outcomes frequently measured in S&C including those measuring maximum strength, jumping ability and sprint performance can be highly correlated (19). In the context of controlled trials, these relationships may result in correlations among comparative effects such that the superiority of an intervention identified in one outcome is related to others measuring the same and potentially different constructs. Where such correlations exist, alpha adjustment approaches that account for these relationships can better balance trade-offs between Type 1 error rates and sample size required for adequate statistical power (20). At present, it is unknown whether comparative effects are correlated in S&C. The purpose of this study, therefore, was to investigate the distribution and potential correlations between comparative effects in S&C controlled trials across typical outcomes and outcome domains. To facilitate description and communication of expected values, thresholds quantifying small, medium and large comparative effects were estimated.

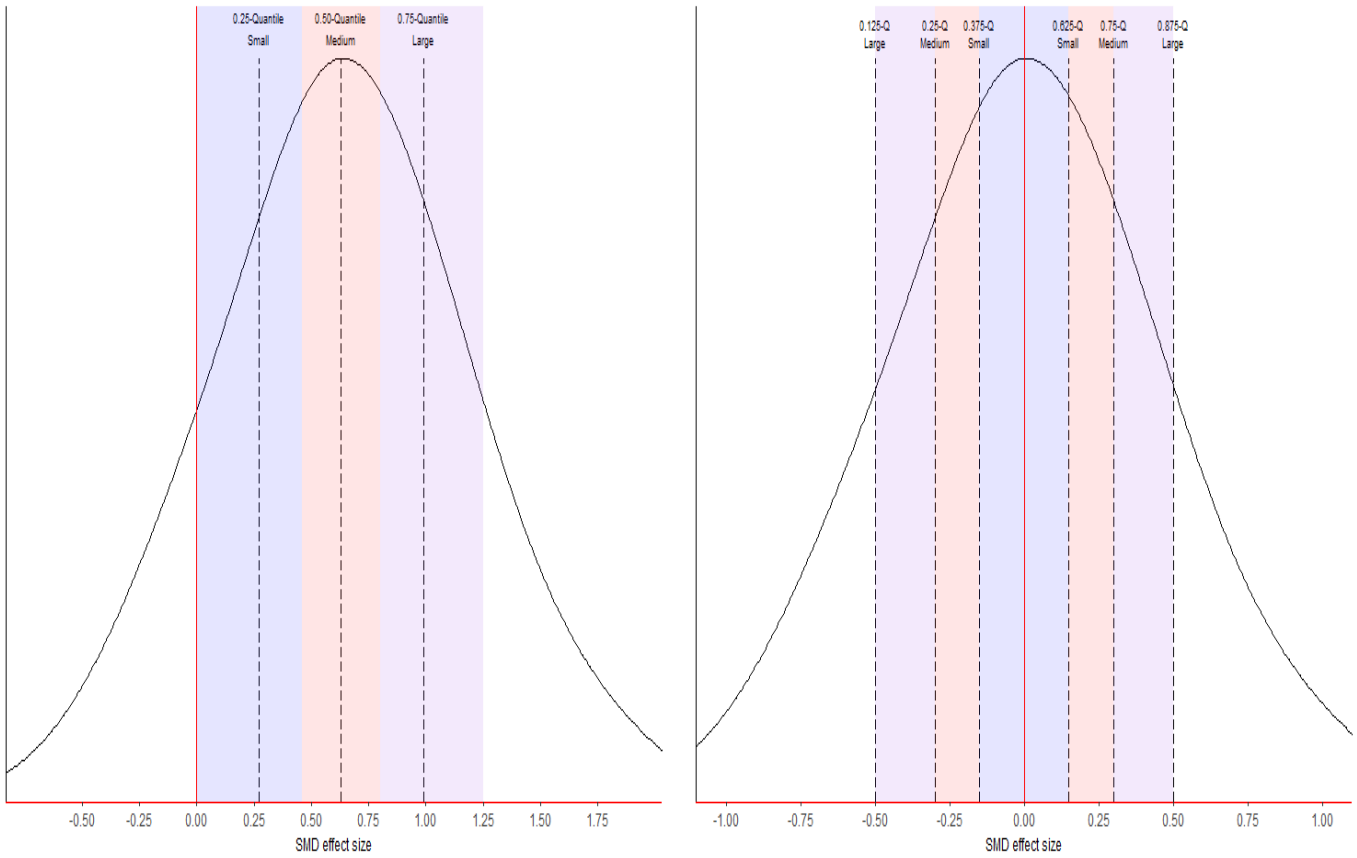
2.0 Methods

2.1 Overview of meta-analysis

The meta-analysis was conducted on a database of S&C training studies obtained from a search of the literature comprising studies from 1962 to 2018. The database included information describing outcome variables along with baseline and follow-up means and standard deviations and has been described elsewhere (1). For the current meta-analysis, outcomes categorised as: 1) maximum strength; 2) power; 3) jump performance; 4) sprint performance; and 5) agility performance were selected. Sub-analyses were also performed on the most frequently measured outcomes to enable calculation of mean difference effect sizes on the original measurement scale and included: 1) 1RM bench press; 2) 1RM squat; 3) unloaded vertical jump height; 4) 10 m sprint time; 5) 20 m sprint time; 6) 30 m sprint time; 7) 40 yard sprint time; and 8) 40

m sprint time. Comparative effect sizes were calculated for studies comprising at least two active interventions (e.g. comparisons did not include no-exercise control or habitual sporting activity only groups). No attempt was made to rank or create a hierarchy for the different interventions or comparisons, but instead, model effect size distributions as they exist across intervention comparisons generally investigated in the S&C literature. The "direction" of the comparative effect sizes was considered random such that across the large database and with additional bootstrapping procedures the effect size distribution would be centred on zero reflecting "sceptical" priors from a Bayesian perspective (21), and two-tailed hypothesis testing from a frequentist perspective. A schematic illustrating the differences between a comparative effect size distribution where the mean is centred on zero, and a non-comparative effect size distribution where most interventions are expected to generate improvements is presented in Figure 1. Previously, the 0.25-, 0.50- and 0.75-quantiles have been used to qualitatively label non-comparative effect sizes as "small", "medium" and "large" (1,22,23). To apply this approach with symmetric comparative effect sizes, the small, medium and large thresholds were defined by the middle 25% (0.375- to 0.625-quantile), the middle 50% (0.25- to 0.75-quantile) and the middle 75% distributions (0.125- to 0.875-quantile), respectively (Figure 1). Following investigation of comparative effect sizes across outcome domains using both standardised and non-standardised statistics, associations between effect sizes were quantified by estimating pairwise correlations.

Figure 1: Schematic illustrating differences in a non-comparative effect size distribution (left) and a comparative effect size distribution centred on zero (right) with small, medium and large thresholds defined.



Q: Quantile; SMD: Standardised mean difference. Most of the non-comparative distribution (left) exceeds zero and regions (small/medium/large) are defined from the left. The comparative distribution (right) is centred on zero with small effects closest to zero in either direction, with medium and large effects located further from the centre.

2.2 Inclusion criteria and data

Inclusion and exclusion criteria for the current meta-analysis were set to include as many relevant S&C training modes and dependent variables as possible. Inclusion criteria comprised: 1) any training-based study ≥ 4 weeks; 2) healthy trained or untrained participants with a mean age between 14 and 60; 3) training group with a minimum of 4 participants; 4) pre- and post-training means and standard deviations; 5) sufficient information provided to appropriately describe the training method; and 6) inclusion of at least two active S&C interventions. Studies comprising training that were predominantly aerobic-based or rehabilitation focused were excluded. Data regarding the study (authors, year, total number of active intervention groups); participant characteristics (final study n, sex, training status, and age); outcome domain (maximum strength, power, jump performance, and sprinting performance); and pre- and post-training means and standard deviations were obtained. The definitions used to categorise outcome domains included: 1) maximum strength: a measure of maximum force production where time was not limited (e.g. 1-6 repetition maximum, isometric mid-thigh pull, peak torque); 2) power: a direct measurement of power output measured in Watts (absolute and normalised relative to body mass); 3) jump performance: measure of jump height or distance; 4) sprint performance: a measurement of the time to complete a specified linear distance or the velocity achieved; and 5) agility performance: a measurement of the time to complete a change of direction or reactive task. Training status was categorised by Rhea et al. (24) based on S&C training experience and categorised as untrained (<1 year), recreationally trained (1-5 years), and highly trained (>5 years). Sex of the groups were categorised as male-only, female-only or mixed sex.

2.3 Statistical analysis

Effect sizes and their sampling variance were calculated using group mean and standard deviation values reported at baseline and at any subsequent time-point. Pairwise comparative standardised mean differences ($\mathrm{SMD}_{AB\mathrm{pre}}$) of an intervention "A" and "B", and their sampling variances σ^2 were calculated using the following formulae (25):

$$\mathrm{SMD}_{AB\mathrm{pre}} = \left(1 - \frac{3}{4(n_A + n_B - 2) - 1}\right)\left(\frac{\left(\bar{x}_{A_{Post}} - \bar{x}_{A_{Baseline}}\right) - \left(\bar{x}_{B_{Post}} - \bar{x}_{B_{Baseline}}\right)}{Sd_{AB_{Pre}}}\right)$$

where n_A and n_B are the number of participants in intervention A and B, the first term comprises a small-study bias term $c(n_A + n_B - 2)$, where $c(n_A + n_B - 2) = 1 - \frac{3}{4(n_A+n_B-2)-1}$, and $Sd_{AB_{Pre}}$ is the baseline pooled standard deviation where $Sd_{AB_{Pre}} = \sqrt{\frac{(n_A-1)Sd_{A_{Pre}}+(n_B-1)Sd_{B_{Pre}}}{n_A+n_B-2}}$.

$$\sigma^2\left(\mathrm{SMD}_{AB\mathrm{pre}}\right) = 2c(n_A + n_B - 2)^2(1-\rho)\left(\frac{n_A + n_B}{n_A n_B}\right)\left(\frac{n_A + n_B - 2}{n_A n_B}\right)\left(1 + \frac{\mathrm{SMD}^2_{AB\mathrm{pre}}}{2(1-\rho)\left(\frac{n_A + n_B}{n_A n_B}\right)}\right) - \mathrm{SMD}^2_{AB\mathrm{pre}}$$

where ρ is the correlation between repeated measures. For sub-analyses conducted on the most common outcomes, mean difference comparative effect sizes were calculated on the original measurement scale and therefore were not standardised by dividing by the baseline pooled standard deviation.

The empirically obtained effect sizes were modelled using a three-level Bayesian mixed effects meta-analytic model. The three levels included the between study (level 3), the outcome (level 2) and the within study sampling variance (level 1). The application of a meta-analytic model enabled sharing of information across studies to better estimate model parameters and accounted for dependencies within the data due to most studies providing more than one data point (based on reporting multiple outcomes and/or multiple time points following baseline) and studies frequently including more than two groups such that multiple pairwise group calculations were made for each outcome. To account for uncertainty in σ^2 due to non-

reporting of correlations between baseline and follow-ups, 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 (26). The parameters obtained from the meta-analysis models were then used to calculate small, medium and large threshold values for each of the outcome domains. This was achieved through bootstrapping and generating posterior predictions. Each analysis included all available studies and data points with 100 bootstrap samples comprising a +1/-1 random allocation for their pairwise effect sizes. The same coefficient was applied to all effect sizes in the study to maintain any associations. For each set of posterior predictions across the bootstrap samples, the 0.625-quantile/|0.375|-quantile, 0.75-quantile/|0.25|-quantile, and 0.875-quantile/|0.125|-quantile values were obtained to quantify small, medium, and large thresholds, respectively. Across the different categories, the median value and spread were used to describe estimates and uncertainty through credible intervals (CrIs).

Correlations were calculated on standardised effect sizes and non-standardised effect sizes for the most common outcomes. To account for dependencies in the data due to single studies providing multiple data points and differences in the precision of estimates, correlations were calculated through three-level weighted meta-regressions with study random effects accounting for systematic differences across studies and sampling errors used to calculate weights. Values were transformed into z-scores so that slope coefficients from meta-regressions were equivalent with correlations ranging from -1 to 1. One-hundred bootstrap samples were applied and uncertainty quantified through summary of the posterior distributions. Default weakly informative Student-t and half Student-t priors with 3 degrees of freedom were used for all intercept and variance parameters, respectively (27). 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 (28). Meta-analyses were performed using the R wrapper package brms interfaced with Stan to perform sampling (29). Convergence of parameter estimates were obtained for all models with Gelman-Rubin R-hat values below 1.1 (30).

.

3.0 Results

3.1 Descriptions of data

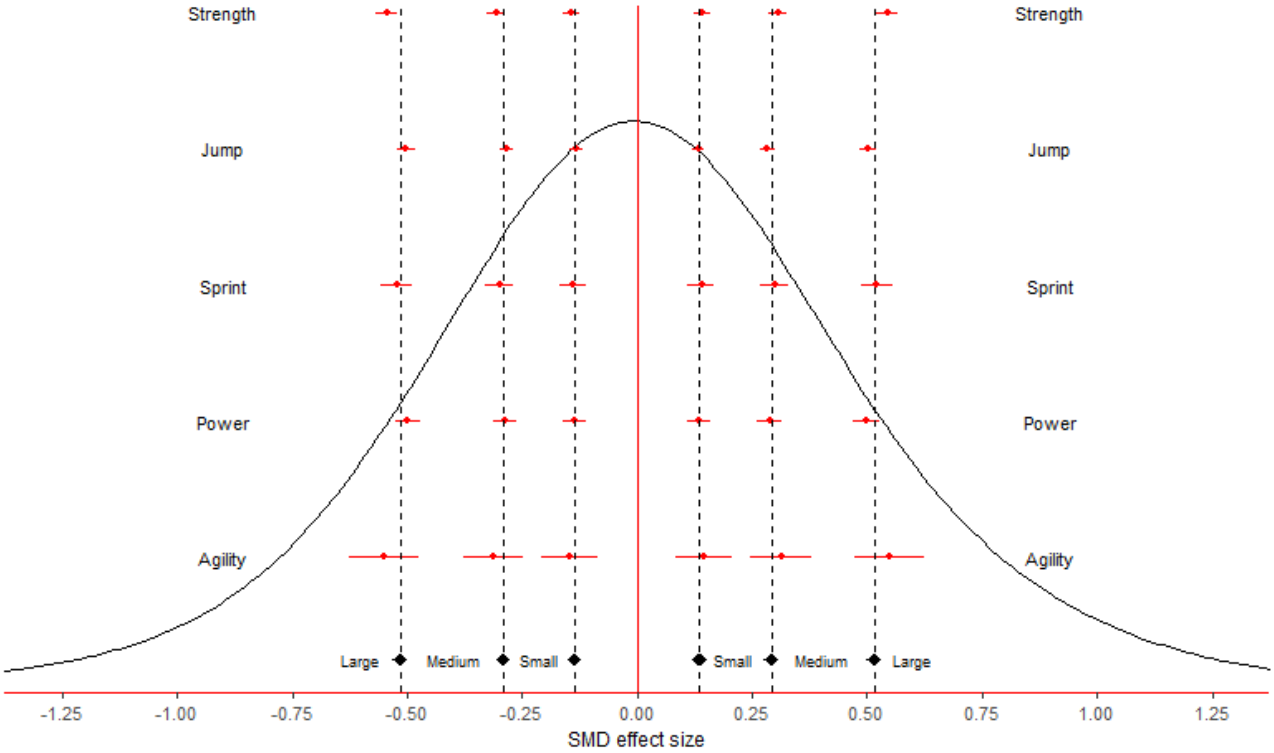
Data to investigate comparative effect sizes were obtained from 417 studies (Supplementary file) that comprised two or more active interventions (318 studies included two groups [76%]; 81 studies included three groups [19%]; 15 studies [4%] included four groups, 2 studies included five groups [0.5%]; and 1 study included 9 groups [0.25%]). Across the 417 studies, 2119 different outcomes (maximum strength: 776 [37%]; jump performance: 453 [21%]; sprint performance: 419 [20%]; power: 355 [17%]; agility performance 116 [5%]) were investigated, which totalled 2430 when considering repetition across multiple time points. A total of 3874 standardised mean difference pairwise effect sizes were extracted. Sub-analyses conducted with the most common outcomes on the initial measurement scale included 1373 mean difference effect sizes (vertical jump height: 527 [40%]; 1RM bench press: 312 [23%]; 1RM squat: 224 [17%]; 20 m sprint time: 106 [8%]; 10 m sprint time: 102 [8%]; 30 m sprint time: 63 [4%]; 40 m sprint time: 20 [2%]; 40 yard sprint time: 19 [1%]).

Across the 417 studies, the median group size was equal to 10 (IQR: 9-13) and the median intervention duration was 8 weeks (IQR: 6-10). Sixty percent of groups were categorised as male-only, 29% were categorised as mixed sex, and 11% were categorised as female-only. Fifty-nine percent of groups were categorised as untrained, 36% were categorised as recreationally trained, and 5% were categorised as highly trained.

3.2 Standardised mean difference comparative effect sizes

A total of 32 outliers were removed from the analysis such that comparative standardised mean difference effects sizes ranged from ±2.7. Application of the meta-analysis model and bootstrapping across all outcomes identified the thresholds as small: $SMD_{AB_{pre}}$= 0.14 [95%CrI: 0.12 to 0.15]; medium: $SMD_{AB_{pre}}$= 0.29 [95%CrI: 0.28 to 0.30]; and large: $SMD_{AB_{pre}}$= 0.51 [95%CrI: 0.50 to 0.53]. No substantive differences in any of the threshold values were identified when the analysis was conducted across the different outcome domains (Figure 2).

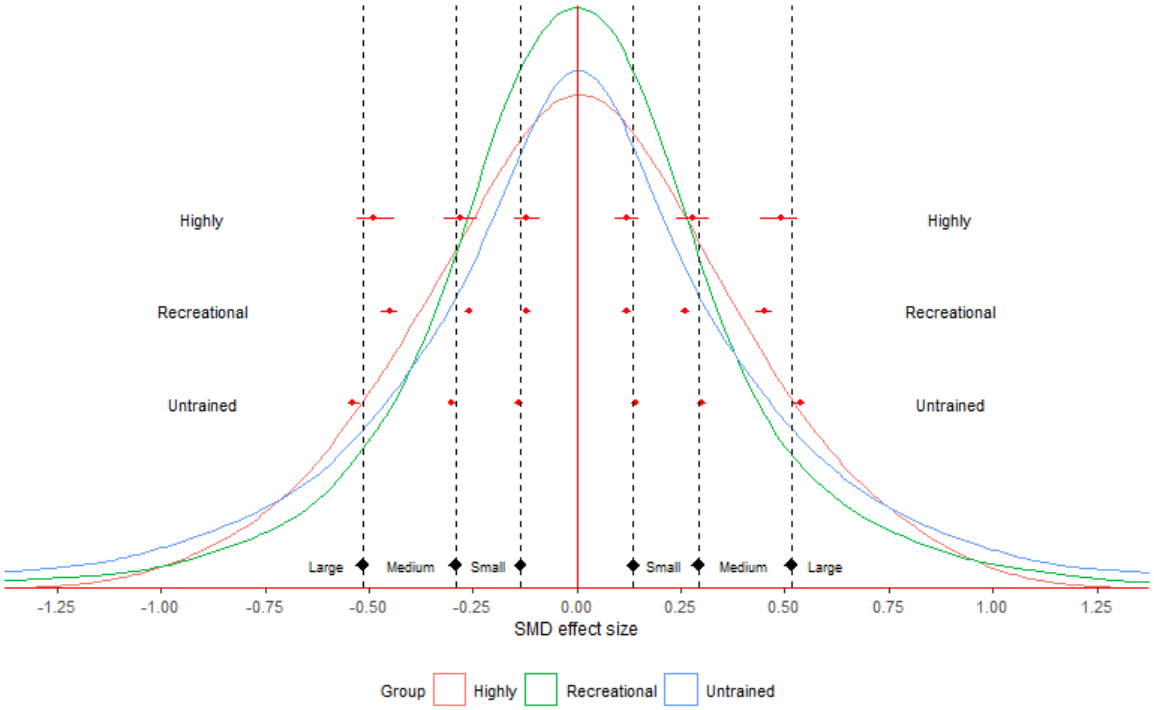
Figure 2: Empirical distribution and modelled comparative effect size thresholds across outcome domains



Black curve is a density plot of the directly calculated empirical comparative effect size values across all outcomes. Small, mid, and large thresholds represent the 0.375-/0.625-, 0.25-/0.75-, and 0.125-/0.875-quantiles of predicted draws. Black diamonds represent threshold values based on all outcomes. Red point ranges illustrate the outcome specific estimates and their uncertainty through the median value (circle) and 95% credible interval.

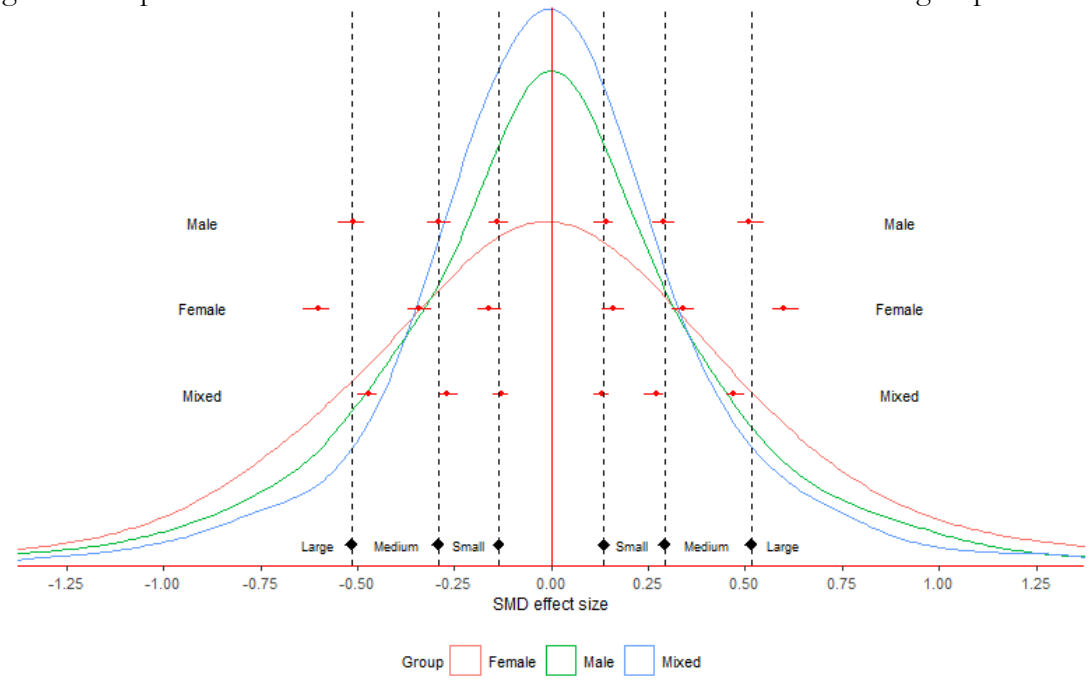
Distributions and thresholds were mainly similar across comparisons with groups of the same training status (untrained, recreationally trained or highly trained; Figure 3) or sex (male-only, female-only or mixed sex; Figure 4). However, potential differences were observed at large thresholds with lower effect sizes for comparisons with recreationally trained groups and greater effect sizes for female-only groups.

Figure 3: Empirical distributions and modelled effect size thresholds across training status groups



Curves are density plots from directly calculated empirical comparative effect size values across all outcomes for comparisons within untrained, recreationally trained, and highly trained groups. Small, medium, and large thresholds represent the 0.375-/0.625-, 0.25-/0.75-, and 0.125-/0.875-quantiles of predicted draws. Black diamonds represent threshold values based on all participants. Red point ranges illustrate the group specific estimates and their uncertainty through the median value (circle) and 95% credible interval.

Figure 4: Empirical distributions and modelled effect size thresholds across groups based on sex



Curves are density plots from directly calculated empirical comparative effect size values across all outcomes for comparisons within male-only, female-only, and mixed sex groups. Small, medium, and large thresholds represent the 0.375-/0.625-, 0.25-/0.75-, and 0.125-/0.875-quantiles of predicted draws. Black diamonds represent threshold values based on all participants. Red point ranges illustrate the group specific estimates and their uncertainty through the median value (circle) and 95% credible interval.

3.2 Mean difference comparative effect sizes

Small, medium, and large thresholds for mean difference comparative effect sizes for commonly measured outcomes are presented in table 1. Substantive overlap was identified in the distributions of the 1RM bench press and 1RM squat, with the greatest divergence found at the large threshold with greater effect sizes obtained with the 1RM squat. Consistent increases in medium and large thresholds were identified across sprint outcomes with greater distance.

3.3 Correlations

Median estimates of correlations between standardised mean difference effect sizes ranged from 0.06 to 0.36, with most values between 0.15 and 0.30. Correlations were higher when they were within an outcome domain (e.g. between strength outcomes) compared with across outcome domains (e.g. between strength and sprint outcomes). Correlations were strongest when they were estimated for non-standardised effect sizes in similar common outcomes (e.g. 1RM bench press with 1RM squat, vertical jump tests, and time to complete, 10, 20 and 30 m sprints) where median estimates ranged from 0.51 to 0.70.

Table 1: Meta-analysis results with small, medium and large thresholds across comparative mean difference effect sizes expressed in the measurement scale of common outcomes.

Outcome	Data	Small [95% CrI]	Medium [95% CrI]	Large [95% CrI]
1RM bench press (kg)	307 effect sizes 87 studies	1.8 [1.1 to 2.6]	4.0 [3.2 to 4.7]	7.1 [6.3 to 7.9]
1RM squat (kg)	214 effect sizes 86 studies	2.7 [1.6 to 3.8]	5.8 [4.7 to 6.9]	10.3 [9.1 to 11.4]
Vertical jump (cm)	521 effect sizes 220 studies	0.7 [0.6 to 0.8]	1.5 [1.3 to 1.6]	2.7 [2.5 to 2.9]
10 m sprint (s)	99 effect sizes 59 studies	0.02 [0.01 to 0.02]	0.03 [0.03 to 0.04]	0.06 [0.05 to 0.07]
20 m sprint (s)	100 effect sizes 62 studies	0.02 [0.01 to 0.02]	0.04 [0.03 to 0.05]	0.07 [0.06 to 0.08]
30 m sprint (s)	60 effect sizes 35 studies	0.03 [0.01 to 0.04]	0.06 [0.04 to 0.08]	0.12 [0.09 to 0.14
40 yard sprint (s)	16 effect sizes 10 studies	0.03 [0.01 to 0.06]	0.07 [0.04 to 0.10]	0.13 [0.09 to 0.17]
40 m sprint (s)	18 effect sizes 14 studies	0.04 [0.00 to 0.08]	0.09 [0.06 to 0.13]	0.16 [0.12 to 0.21]

1RM: One repetition maximum; CrI: Credible interval. Small threshold is the |0.375|-/0.625-quantile; Medium threshold is the |0.25|-/0.75-quantile; Large threshold is the |0.125|-/0.875-quantile.

4.0 Discussion

The primary aim of this study and incorporated meta-analyses was to investigate comparative effect sizes distributions across S&C studies comparing active interventions. The analyses identified that comparative standardised mean difference effect sizes in S&C are generally low in magnitude with the middle 50% of the distribution ranging between approximately ±0.3. Analyses also found that the majority of comparative

effect distributions are likely to be similar across outcome domains and different participants groups in terms of training status and sex. Within frequentist frameworks, these finding have implications for powering of future studies and the need for substantively larger sample sizes than have been used previously when conducting standard pre-post parallel group controlled designs. The secondary aim was to investigate correlations between comparative effects such that any superiority in one intervention may be reflected across multiple outcomes. Correlations ranged widely in magnitude, but were positive and in general stronger among outcomes within the same outcome domain (e.g. between strength measures) and highest among similar outcomes when expressed in absolute magnitude (e.g. between 1RM bench press and 1RM squat). These findings indicate that where there are differences among interventions in terms of average treatment effect, these differences are likely to be observable across multiple outcomes, influencing approaches researchers should adopt when choosing to control for multiple tests to ensure statistical power is not unduly lowered.

Controlled studies in S&C have explored an extensive space of possible training approaches. Frequently, comparisons have been made with popular training interventions with relatively minor adjustments, for example, the use of variable resistance (e.g. elastic bands or chains) in comparison to traditional resistance training (31), or alterations to sequence of exercises (e.g. complex vs contrast training) (32). These relatively minor adjustments reflect a desire to try and optimise training responses, however, it may therefore not be surprising that the bulk of comparative effect sizes in S&C should be expected to generate values that are low in magnitude and close to zero. Describing the middle 25, 50, and 75% of the modelled distribution as small, medium and large, estimates of approximately 0.15, 0.30, and 0.50 were obtained, respectively. A priori sample size calculations employing a frequentist framework and a simple statistical analysis (two-tailed independent t-test with change scores) to test the null hypothesis of zero population average treatment effect, with α=0.05, and power (1-β) = 0.80, returns group sizes of 699, 176, and 64 for the different threshold values (14). Similar sample size calculations are achieved when considering the use of a repeated measures ANOVA (between-within design) and the interaction effect to test for differences in average treatment effect, as is common in S&C research. Here, the pattern of means, standard deviations and correlation between pre- and post-interventions scores influence calculations (13). Using the original measurement scale and data obtained for the 1RM squat, the estimated small, medium and large improvements of 3, 6, and 10 kg obtained from this study identify the same required group sizes of 699, 176, and 64 when assuming a correlation of 0.5 (combined with baseline mean of 110 kg and standard deviation of 20 kg) (13). Increasing the correlation to 0.7 lowers the sample size requirements to 420, 106, and 39 given the same threshold values (13). Across the 417 studies and 958 groups used to obtain data for this study, the median group size was equal to 10 (IQR: 9-13), with 8 studies including groups sizes greater than 50 and the maximum equal to 94. These differences highlight how underpowered S&C research has tended to be and the challenge of adequately powering future research.

In our previous meta-analysis of non-comparative effect sizes across the S&C literature (1), substantive differences were observed in distributions across outcome domains, with the greatest difference identified between those measuring maximum strength and sprint performance. In the present study these differences were not observed, indicating that whilst the magnitude of change caused by any single intervention is likely to be different across outcome domains, the relative difference between two interventions is not. This finding has important implications for future research in terms of sample sizes and the knowledge that the ability to make inferences will be similar for different outcomes and for example many more participants will not be required for research investigating the development of speed compared to the development of maximum strength. The presentation of comparative effect sizes in the original measurement scale for the common outcomes of the 1RM squat, 1RM bench press, vertical jump height and sprint times (Table 1) provides researchers with additional information that may be used to perform

sample size calculations across a broader array of trial designs using methods previously highlighted for S&C (33) and more general disciplines (13).

Analyses investigating potential differences in comparative effect size distributions across participant groups in terms of training status and sex identified that the bulk of the distributions were similar, however, differences may exist towards the large threshold. Results identified that the large threshold was lower for recreationally trained individuals, and greater for female-only groups. These potential differences align with findings from previous research showing greater relative improvements in outcomes such as strength for untrained participants and females (34,35). It has been hypothesised that such differences may be due to greater capacity to improve based on a general lower starting point (35). Greater confidence in these findings would have been obtained if ordered effects were observed such that the effect size distribution was narrower for highly trained participants (untrained > recreational > highly), and effects for mixed sex groups were between male- and female-only groups. Such ordered effects were not found (Figures 3 & 4) and may be due to several reasons. For training status, only 5% of groups included in the data comprised highly trained participants, therefore descriptions of the effect size distribution for this population were more uncertain and considerable overlaps between estimates for highly and recreationally trained groups were identified. For mixed sex groups, the percentage of males and females was generally not equal and varied considerably across studies, such that the category is potentially limited in its capacity to act as an intermediate in any comparison.

Results from our previous meta-analyses have shown that most studies investigating interventions in S&C measure multiple outcomes, often across multiple time points with more than two groups (34). On average, studies were shown to include 13 different data points that could be used to test for average treatment effects (34). Assuming the null-hypothesis may be true in many of these cases, this level of multiple testing has the potential to substantively increase the Type I error rate where constituent null hypotheses are subjected to a disjunction testing approach. Whilst it may be rare for researchers to explicitly state joint hypotheses that are being tested with multiple outcomes and the specific testing approach adopted (e.g. disjunction, conjunction or individual) (15), the frequent use of multiple outcomes that measure the same domain and conclusion that one intervention is superior to another when some outcomes fail to reach statistical significance indicates a general use of disjunction testing and potential increased Type I error rate. Given the observation that statistical power may already be low in S&C research given low effects and the use of small samples, alpha adjustment for multiple testing will further reduce statistical power causing additional challenges (15). It is therefore important that any procedure adopted is not unduly conservative. Previous simulations have shown that alpha adjustment approaches such as Dubey/Armitage-Parmar (36) that account for correlations between outcomes outperform popular methods such as Bonferroni when correlations exceed 0.3, and that these improvements increase with the adoption of more outcomes (20). In the present study, correlations between comparative effect sizes were shown to vary widely, but were consistently higher between outcomes measuring the same domain. Median estimates were highest for comparative standardised mean differences from outcomes measuring power ($r = 0.36$ [95%CrI: 0.28 to 0.43]) and jump performance ($r = 0.33$ [95%CrI: 0.24 to 0.41]). Sub-analyses conducted on the most common outcomes measured in their original scale returned median estimates ranging between 0.51 and 0.70. Given the potentially strong correlations that exist between comparative effect sizes it is clear that any alpha adjustment approach to protect against issues with multiplicity should account for these associations. Future simulation work creating data reflective of S&C interventions is required to best understand how to balance Type I and II errors and the genuine interest of researchers testing interventions across multiple outcome domains.

The usefulness of labelling effect sizes as small, medium, and large has been questioned in S&C (33) as well as other disciplines (37). Much of the criticism has surrounded the arbitrariness of the original thresholds proposed by Cohen (38) and the intuition that these are likely to be different in specific contexts within a discipline (33). It has been argued that effect-sizes should be interpreted based relative to their costs (i.e. practical or substantive significance), other effects in the same empirical context (as is provided in the present study), or using benchmarks such as the smallest effect of interest (37). The results from the present study highlight that given the intervention comparisons that have generally been studied in S&C, comparative effects are low in magnitude, thereby requiring sample sizes for standard control designs and analyses that have not typically been used. If benchmarks such as smallest effect of interest do not coincide with the values presented here, it is likely that the values are too large and therefore unlikely that a reasonable comparison is being made (e.g. use of an inappropriate control such as a non-active intervention), or values are too low such that sample sizes required would not be feasible. It is important to note, however, that the majority of intervention studies in S&C are extremely short in duration with most data points measuring change after six to twelve weeks (1). As a result, larger comparative effect sizes may occur over longer durations when differences between interventions have a greater change to manifest.

It has previously been suggested that small effect sizes that may be relatively common in S&C and sport science in general, may benefit from Bayesian analyses to improve estimation and provide a means of presenting results that are straightforward to interpret in terms of probabilistic statements (39). It has also been stressed, however, that Bayesian methods do not represent a panacea that can overcome challenges such as sample sizes that are too small or questionable research practices (10). Given the extensive research that has been conducted in S&C, the results obtained from analyses such as those presented in the current study may be useful in developing informative priors to improve estimates. For example, the primary meta-analysis presented here across all standardised mean difference effect sizes generated a distribution centred on zero with small, medium and large thresholds equal to approximately 0.15, 0.30 and 0.50, respectively. These values and the overall distribution can be reasonably modelled with a Gaussian distribution with mean equal to 0 and standard deviation equal to 0.46 (0.625-, 0.725-, and 0.875-quantiles equal to 0.15, 0.31, and 0.53). Where this empirically derived distribution aligns with a researchers' beliefs, it could be used as a prior distribution. Bayesian updating using methods outlined by Jones et al. (40) and previously applied in S&C contexts (26), could then be combined with incoming data to generate posterior distributions using simple formulas and enabling a range of probabilistic interpretations to be made (26). The distribution created in the present study represents a so-called sceptical prior centred on zero. Where there is more confidence that an intervention is superior to the reference, alternative priors could be used including similar distribution in terms of spread (e.g. similar standard deviation) that are shifted to be centred on the small, medium or large thresholds. Additionally, different prior distributions could be used to account for specifics in the researchers' beliefs. This could include use of t-distributions to reflect a view that larger effect size magnitudes may occur relatively frequently, or skew distributions reflecting the belief that larger positive results may be plausible but larger negative results favouring the standard are not. Incorporation of these more complex priors will require more sophisticated analysis methods to generate posterior distributions and may benefit from further meta-analysis work or prior elicitation with experts (41).

5.0 Practical Applications

The use, analysis and interpretation of effect sizes is complex in all areas of research including S&C. A single approach with effect sizes is unlikely to be optimum in all contexts and limitations including the appropriateness of underlying assumptions presents challenges. In the present study, an approach consistent with development of sceptical priors and two-sided null-hypothesis testing was developed to summarise comparative effect sizes in S&C. The results indicated that most comparative effect sizes should

be expected to be low in magnitude, such that substantive resources including large sample sizes would be required to reliably estimate the population average treatment effect. If it is deemed relevant to try and estimate comparative standardised mean difference effect sizes potentially as low as 0.15, approaches different from those used previously will be required. Principally, larger samples are required which may be achieved through collaborating research teams conducting the same protocol across different sites. Additionally, focus may be placed on increasing the duration of intervention studies under the assumption that comparative effects will increase in magnitude and be easier to identify. However, issues of attrition and heterogenous responses may become more impactful over long duration studies. Researchers may also seek to use and develop different analysis techniques that leverage high frequency data collection to enhance the amount and structure of the information to increase the precision of estimates. Further statistical work including detailed simulations may present a cost-effective means to compare these different strategies incorporating the information presented in this study to fit the S&C context and provide needed future guidance.

Acknowledgements

No funding was received for this review.

Conflicts of interest

Paul Swinton and Andrew Murphy declare that they have no potential conflicts of interest with the content of this article.

References

(1) Swinton PA, Burgess K, Hall A, Greig L, Psyllas J, Aspe R, et al. Interpreting magnitude of change in strength and conditioning: Effect size selection, threshold values and Bayesian updating. *Journal of Sports Sciences.* 2022; In Press.

(2) Rhea MR, Alvar BA, Burkett LN. Single versus multiple sets for strength: a meta-analysis to address the controversy. *Research Quarterly for Exercise and Sport.* 2002; 73:485-488. https://doi.org/10.1080/02701367.2002.10609050

(3) Ralston GW, Kilgore, L. Wyatt, F.B. Baker, J.S. The effect of weekly set volume on strength gain: A meta-analysis. *Sports Medicine.* 2017; 47:2585-2601. https://doi.org/10.1007/s40279-017-0762-7

(4) Williams TD, Tolusso DV, Fedewa MV, Esco MR. Comparison of periodized and non-periodized resistance training on maximal strength: A meta-analysis. *Sports Medicine.* 2017; 47:2083-2100. https://doi.org/10.1007/s40279-017-0734-y

(5) Harries SK, Lubans DR, Callister R. Systematic review and meta-analysis of linear and undulating periodized resistance training programs on muscular strength. *The Journal of Strength and Conditioning Research.* 2015; 29:1113-1125. https://doi.org/10.1519/JSC.0000000000000712

(6) Schoenfeld BJ, Wilson JM, Lowery RP, Krieger JW. Muscular adaptations in low-versus high-load resistance training: A meta-analysis. *European Journal of Sport Science.* 2016; 16:1-10. https://doi.org/10.1080/17461391.2014.989922

(7) Morris SJ, Oliver JL, Pedley JSH, G.G., Lloyd RS. Comparison of weightlifting, traditional resistance training and plyometrics on strength, power and speed: a systematic review with meta-analysis. *Sports Medicine.* 2022; 13:1-22. https://doi.org/10.1007/s40279-021-01627-2

(8) Maier M, Lakens D. Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices Psychological Science.* 2022; 5:1-14. https://doi.org/10.1177/25152459221080396

(9) Cohen J. The earth is round (p<.05). *American Psychologist.* 1994; 49:997-1003. https://psycnet.apa.org/doi/10.1037/0003-066X.49.12.997

(10) Lohse K. No Estimation without Inference: A Response to the International Society of Physiotherapy Journal Editors. *Communications in Kinesiology.* 1(4). https://doi.org/10.51224/cik.2022.49

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

(12) Angus DC, Chang CCH. Heterogeneity of treatment effect. Estimating how the effects of interventions vary across individuals. *JAMA.* 2021; 326(22):2312-2313. https://doi.org/10.1001/jama.2021.20552

(13) Lakens DC, A. Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science.* 4. https://doi.org/10.1177/2515245920951503

(14) Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods.* 2009; 41:1149-1160. https://doi.org/10.3758/BRM.41.4.1149

(15) Rubin M. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. *Synthese.* 2021; 199:10969-11000. https://doi.org/10.1007/s11229-021-03276-4

(16) Midway S, Robertson M, Flinn S, Kaller M. Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. *PeerJ.* 2020; 8:e10387. https://doi.org/10.7717/peerj.10387

(17) Sinclair JK, Taylor PJ, Hobbs SJ. Alpha level adjustments for multiple dependent variable analyses and their applicability. *International Journal of Sports Science and Engineering.* 2013; 7:17-20.

(18) VanderWeele TJ, Mathur MB. Some desirable properties of the Bonferroni correction: Is the Bonferroni correction really so bad? *American Journal of Epidemiology.* 2019; 188:617-618. https://doi.org/10.1093/aje/kwy250

(19) Swinton PA, Lloyd R, Keogh JWL, Agouris I, Stewart AD. Regression models of sprint, vertical jump, and change of direction performance. *Journal of Strength and Conditioning Research.* 2014; 28:1839-1848. https://doi.org/10.1519/JSC.0000000000000348

(20) Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC Medical Research Methodology.* 2019; 19:129. https://doi.org/10.1186/s12874-019-0754-4

(21) Zampieri FG, Casey JD, Shankar-Hari M, Harrell FE, Harhay MO. Using Bayesian methods to augment the interpretation of critical care trials. An overview of theory and example reanalysis of the alveolar recruitment for acute respiratory distress syndrome trial. *American journal of respiratory and critical care medicine.* 2021; 203:543-552. https://doi.org/10.1164/rccm.202006-2381CP

(22) Brydges CR. Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innovation in Aging.* 2019; 3:igz036. https://doi.org/10.1093/geroni/igz036

(23) Gignac GE, Szodorai ET. Effect size guidelines for individual differences researchers. *Personality and Individual Differences.* 2016; 102:74-78. https://doi.org/10.1016/j.paid.2016.06.069

(24) Rhea M. Determining the magnitude of treatment effects in strength training research through the use of effect sizes. *Journal of Strength and Conditioning Research.* 2004; 18:918-920. https://doi.org/10.1519/14403.1

(25) Morris SB. Estimating effect sizes from pretest-posttest-control group design. *Organizational Research Methods.* 2008; 11:364-386. https://doi.org/10.1177/1094428106291059

(26) Swinton PA, Burges K, Hall A, Greig L, Psyllas J, Aspe R, et al. 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.* 2021. https://doi.org/10.51224/SRXIV.11

(27) Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis.* 2006; 1:515-534. https://doi.org/10.1214/06-BA117A

(28) Verardi V, Vermandele C. Univariate and multivariate outlier identification for skewed or heavy-tailed distributions. *The Stata Journal.* 2018; 18:517-532. https://doi.org/10.1177/1536867X1801800303

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

(30) Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Taylor & Francis; 2014. https://doi.org/10.1201/b16018

(31) Soria-Gila MA, Chirosa IJ, Bautista IJ, Baena S, Chirosa LJ. Effects of variable resistance training on maximal strength: A meta-analysis. *Journal of Strength and Conditioning Research.* 2015; 29:3260-3270. https://doi.org/10.1519/JSC.0000000000000971

(32) Cormier P, Freitas TT, Rubio-Arias JÁ, Alcaraz PE. Complex and Contrast Training: Does Strength and Power Training Sequence Affect Performance-Based Adaptations in Team Sports? A Systematic Review and Meta-analysis. *Journal of Strength and Conditioning Research.* 2020; 34:1461-1479. https://doi.org/10.1519/JSC.0000000000003493

(33) Beck TW. The importance of a priori sample size estimation in strength and conditioning research. *Journal of Strength and Conditioning Research.* 2013; 27:2323-2337. https://doi.org/10.1519/JSC.0b013e318278eea0

(34) Swinton PA, Burgess K, Hall A, Greig L, Psyllas J, Aspe R, et al. A Bayesian approach to interpreting intervention effectiveness in strength and conditioning: Part 1. A meta-analysis to derive context-specific thresholds. *Pre-print available from SportRχiv.* 2021. https://doi.org/10.51224/SRXIV.9

(35) Roberts BM, Nuckols G, Krieger JW. Sex differences in resistance training: A systematic review and meta-analysis. *Journal of Strength and Conditioning Research.* 2020; 34:1448-1460. https://doi.org/10.1519/JSC.0000000000003521

(36) Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine.* 1997; 16:2529-2542.

https://doi.org/10.1002/(sici)1097-0258(19971130)16:22<2529::aid-sim692>3.0.co;2-j

(37) Primbs MA, Pennington CR, Lakens D, Silan MAA, Lieck DSN, Forscher PS, et al. Are Small Effects the Indispensable Foundation for a Cumulative Psychological Science? A Reply to Götz et al. (2022). *Perspectives on Psychological Science.* 2022. https://doi.org/10.1177/17456916221100420

(38) Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd. ed. Hillsdale, NJ: Lawrence Erlbaum Associate; 1988. https://doi.org/10.4324/9780203771587

(39) Mengersen KL, Drovandi CC, Robert CP, Pyne DP, Gore CJ. Bayesian estimation of small effects in exercise and sports science. *PLos ONE*. 2016; 11:e014731. https://doi.org/10.1371/journal.pone.0147311

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

(41) Stefan AM, Evans NJ, Wagenmakers EJ. Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*. 2022; 27:177-197. https://doi.org/10.1037/met0000354