

Meta-Analysis of Variation in Sport and Exercise Science*

Examples of Application Within Resistance Training Research

James Steele^a

James Fisher^a

Dave Smith^b

Brad Schoenfeld^c

Yefeng Yang^d

Shinichi Nakagawa^d

12th September 2023

Abstract

Meta-analysis has become commonplace within sport and exercise science for synthesising and summarising empirical studies. However, most research in the field focuses upon mean effects; particularly the effects of interventions to improve outcomes such as fitness or performance. It is thought that individual responses to interventions vary considerably. Hence, interest has increased in exploring *precision* or *personalised* exercise approaches. Not only is the mean often affected by interventions, but variances may also be impacted. Exploration of variances in studies such as randomised controlled trials (RCTs) can yield insight into interindividual heterogeneity in response to interventions and help determine generalisability of effects. Yet, larger samples sizes than those used for typical mean effects are required when probing variances. Thus, in a field with small samples such as sport and exercise science, exploration of variance through a meta-analytic framework is appealing. Despite the value of embracing and exploring variation alongside mean effects in sport and exercise science it is rarely applied to research synthesis through meta-analysis. We introduce and evaluate different effect size calculations along with models for meta-analysis of variation using relatable examples from resistance training RCTs.

1 Introduction

Although the quantitative synthesis of results across studies has existed since the 17th century (Plackett, 1958), the modern-day term “meta-analysis” was coined by Gene Glass (1976). Since that time, the use of meta-analysis as a tool for the synthesis of research in sport and exercise science has increased considerably (Hagger, 2022), with resistance training (RT) accounting for a considerable proportion of this growth (figure 1). Accordingly, throughout the paper we use RT studies as a hopefully familiar example for sport and exercise science researchers. However, to begin we provide a conceptual overview of meta-analyses and effect sizes as typically used within sport and exercise science.

There are two popular models¹ used for meta-analysis: the fixed-effect model and the random-effects model (Borenstein et al., 2010). The fixed effect model assumes that there is one true effect size² that each study included in the meta-analysis has estimated and that any differences in the estimates between individual

*Preprint, please cite as: Steele, J., Fisher, J., Smith, D., Schoenfeld, B., Yang, R., & Nakagawa, S. (2022). Meta-Analysis of Variation in Sport and Exercise Science: Examples of Application Within Resistance Training Research. *sporrxiv* DOI: <https://doi.org/10.51224/SRXIV.214>. Address for correspondence: james.steele@solent.ac.uk Affiliations: ^aFaculty of Sport, Health, and Social Sciences, Solent University, UK; ^bResearch Centre for Musculoskeletal and Sports Medicine, Manchester Metropolitan University, Manchester, UK; ^cHealth Sciences Department, CUNY Lehman College, Bronx, New York, USA; ^dEvolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia

¹To clarify language here for those unfamiliar, the term and concept *model* is used commonly in statistics. A statistical model essentially is specification of what we think the data generating process might be for a given situation. In the context of meta-analyses the data are usually the individual effects that we have extracted from studies i.e., the results of each study. The model, in mathematical formulae, is intended to approximate the processes that we assume led to the generation of the data.

²*Effect size* is an agnostic term used for a family of statistics which communicate the strength of a given ‘effect’ resulting from research. This includes descriptive statistics ranging from mean raw values to correlation coefficients and everything in between (Caldwell & Vigotsky, 2020) including, as we shall see, statistics describing variation.

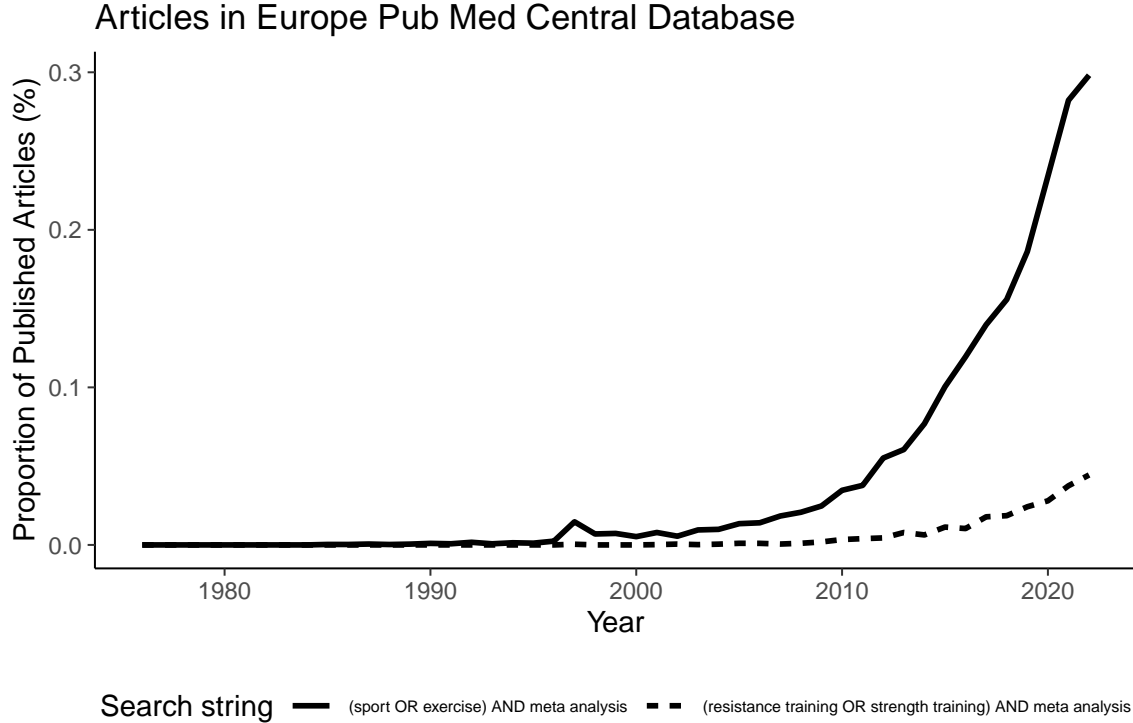


Figure 1: Trends in meta-analyses published in sport and exercise science since 1976.

studies are due to only sampling error. This essentially means that there is a single common effect which is fixed across studies and each study takes samples of individuals from the population to estimate this effect. We can express this model in the following formula:

$$\hat{\theta}_i = \theta + m_i \tag{1}$$

where $\hat{\theta}_i$ is the i th effect size ($i = 1, 2, \dots, N_i$; where N_i is the number of studies and thus effect sizes), θ is the intercept or overall mean (i.e., the fixed-effect), and m_i is the sampling error for each effect size normally distributed with $\sigma_{m_i}^2$. Studies with smaller standard errors of their effect estimates have smaller sampling errors and so these higher precision estimates are given greater weight in the model. The weighting given to each study is calculated as:

$$w_i = \frac{1}{s_i^2} \tag{2}$$

Where s_i is the standard deviation for the effect estimate and thus s_i^2 is the variance. It is referred to as inverse-variance weighting. Then, the overall weighted mean effect estimate from the model is then calculated as:

$$\hat{\theta} = \frac{\sum_{i=1}^I \hat{\theta}_i w_i}{\sum_{i=1}^I w_i} \tag{3}$$

Contrastingly, the random effects model does not make the assumption that there is only one fixed-effect. Instead it allows for the true effects that each study estimates to differ. Each study may share a common underlying effect size but, due to differences between studies in factors such as population characteristics,

the manner in which outcomes are operationalised, or subtle differences in intervention and context to name a few, it is possible that the actual effect being estimated by each study differs. The assumption of the random-effects model then is that the studies included estimate effects that come from a larger population of effects determined by the inclusion criteria for studies included. So, the model assumes that the studies included reflect a random sample of all possible permutations of study and the effects they estimate from this population distribution of studies and effects. Hence, in the fixed-effect model there is one effect and it is assumed to be *fixed* across studies, whereas in the random-effects model there are many and we examine an assumed *random* sample of them. We can express this model in the following formula:

$$\hat{\theta}_i = \theta_i + \tau_i + m_i \quad (4)$$

where $\hat{\theta}_i$ is the i th effect size ($i = 1, 2, \dots, N_i$; where N_i is the number of studies and thus effect sizes), θ_i is the intercept or overall mean of the effects across each study, τ_i is the deviation from θ_i for the i th study (i.e., the random-effects), and m_i is the sampling error for each effect size normally distributed with $\sigma_{m_i}^2$. Essentially this model assumes that each individual study estimates an effect and there will be some sampling error in estimating it. But, the effects each study estimates comes from an overall distribution of true effects with a mean value (i.e., θ_i). The weighting in any model employing random effects such as this requires a different approach as the variance of the distribution of the effect sizes the model assumes, known as τ^2 , must also be estimated³. This essentially describes the heterogeneity between the effects included. Once this has been estimated we can calculate a weight that is adjusted for the random effects for each effect as:

$$w_i^* = \frac{1}{s_i^2 + \tau^2} \quad (5)$$

Then the average of the distribution of effects can be calculated as per equation (3) substituting w_i for w_i^* .

Although historically fixed effects models were commonplace in the field of sport and exercise science (Hagger, 2006) nowadays the random effects model is more often employed (Hagger, 2022). This is likely due to the fact that direct (i.e., exact) replication of studies is rare⁴ and instead studies often explore similar effects across varying moderating factors such as those noted above.

As with many other fields (Mills et al., 2021; Nakagawa et al., 2015; Usui et al., 2021) likely the most common aim in meta-analysis in sport and exercise science, and indeed primary empirical research too, is to estimate the *effect* of an independent variable upon some dependent outcome variable. The dependent variable is often the mean of a measurement and the independent variable is often a categorical grouping; for example, the comparison of an intervention group(s) and a control group, the comparison of intervention groups between one another, or comparison between non-manipulated categories such as biological sex. Indeed, a recent umbrella review (Bernárdez-Vázquez et al., 2022) of meta-analyses in RT identified 14 studies examining the manipulation of RT intervention variables (i.e., the comparison of one intervention to another whereby a variable in the intervention was manipulated) on hypertrophy outcomes, all of which focused on the comparison of mean changes between different intervention groups.

Often, due to the varying operationalisations used for broad outcome concepts⁵, an effect size is used which is “standardised” across studies to allow for their synthesis. Most commonly, a magnitude based⁶ effect size statistic (Caldwell & Vigotsky, 2020), the standardised mean difference (SMD), is used to compare means between groups or conditions. This statistic is usually a version of Cohen’s d (Cohen, 1988), or its

³This estimation can be done using a variety of methods and is an area of ongoing investigation as to how different methods perform. This is beyond the scope of this paper to discuss. We note however that the models we present all utilise Restricted Maximum Likelihood estimation.

⁴Hence current efforts to conduct direct replications (see <https://ssreplicationcentre.com/>).

⁵For example, *strength* might be examined in different studies using different operationalisations including one repetition maximum testing or maximum voluntary contractions. Or the same operationalisations may be employed but different exercises such as the squat or bench press.

⁶Though notably not all meta-analyses use *magnitude based* effect sizes. Indeed some explicitly use what Caldwell and Vigotsky (2020) term *signal-to-noise* effect sizes (e.g., Heidele et al. (2022)).

bias-corrected⁷ metric referred to as Hedges' g [Hedges & Olkin (2014); Borenstein et al. (2021); Nakagawa & Cuthill (2007)]⁸. The SMD, and its sampling variance, s_{SMD}^2 are given by:

$$SMD = \frac{\bar{x}_E - \bar{x}_C}{s_{pooled}} J \quad (6)$$

$$J = 1 - \frac{3}{4(n_C + n_E) - 2} \quad (7)$$

$$s_{pooled} = \sqrt{\frac{(n_C - 1)s_C^2 + (n_E - 1)s_E^2}{n_C + n_E - 2}} \quad (8)$$

$$s_{SMD}^2 = \frac{n_C + n_E}{n_C n_E} + \frac{SMD^2}{2(n_E + n_C)} \quad (9)$$

where \bar{x}_C and \bar{x}_E are the sample means of the control group (C) and experimental (E) or intervention group respectively, s_C and s_E are the standard deviations of the two groups, n_C and n_E are the sample sizes of the two groups, and J is a bias correction for small sample sizes.

The natural logarithm of the ratio of two means ($\ln RR$) is also another effect size statistic that can be used (Curtis & Wang, 1998; Hedges et al., 1999; Lajeunesse, 2011, 2015). The $\ln RR$, and its sampling variance, $s_{\ln RR}^2$ are given by:

$$\ln RR = \ln \frac{\bar{x}_E}{\bar{x}_C} \quad (10)$$

$$s_{\ln RR}^2 = \frac{s_C^2}{n_C \bar{x}_C^2} + \frac{s_C^4}{2n_C \bar{x}_C^2 \bar{x}_C^4} + \frac{s_E^2}{n_E \bar{x}_E^2} + \frac{s_E^4}{2n_E \bar{x}_E^2 \bar{x}_E^4} \quad (11)$$

Due to its calculation the SMD is affected not only by the difference in means of the two groups, but also by the standard deviations of both groups due to the standardisation of the effect size by s_{pooled} in the denominator. In contrast, the $\ln RR$ is uninfluenced by the standard deviations in either groups (see equation (10)), which only affects the sampling variance (see equation (11)). Despite this, the use of effect sizes like the $\ln RR$ has been limited in previous meta-analyses in sport and exercise science (Deb et al., 2018; Nuzzo et al., 2023a) and to our knowledge only a couple of meta-analyses of RT interventions has used this kind of effect size (Swinton et al., 2022; Wolf et al., 2023).

Although researchers in sport and exercise science, among other fields, have focused on estimating the average effects of interventions using randomised trial designs for both primary research and synthesis through meta-analysis, responses to certain interventions may vary on a subgroup or even individual basis. The increased interest in *precision* or *personalised* approaches to exercise prescription has resulted in a number of opinion and methodological review articles discussing statistical approaches to understanding interindividual response heterogeneity to exercise interventions (Atkinson et al., 2019; Atkinson & Batterham, 2015; Hecksteden et al., 2015; Hopkins, 2015; Hrubeniuk et al., 2022; Kelley, 2022; Pickering & Kiely, 2019; Ross et al., 2019; Swinton et al., 2018). However, despite the availability of approaches to compare variances between groups, in sport and exercise science this is rarely explored in primary research (Bonafiglia et al., 2022). Moreover, although there has been increased interest in recent years, few meta-analyses in sport and exercise include both

⁷For those unfamiliar with the terminology, an estimator for a statistic is unbiased if it produces parameter estimates that are on average correct. Thus a bias corrected statistic is one which would be biased without the correction applied, but otherwise has been shown to be unbiased.

⁸We will refer to both merely as the SMD throughout the manuscript for simplicity and note that throughout when reporting a 'SMD' we are reporting the bias-corrected version. We also note that another magnitude based effect size, Glass' Δ , is commonly recommended as it is the simplest form of SMD though makes assumptions about the impact of the intervention having no effect on the denominator (i.e., variance; Caldwell & Vigotsky (2020)).

comparisons of means and variances or explicitly aim to investigate the latter (Bonafiglia et al., 2022; Esteves et al., 2021; Fisher et al., 2022; Kelley et al., 2020, 2022; Steele et al., 2021). Examination of interindividual heterogeneity in response to interventions presents considerable value to researchers and practitioners in sport and exercise science; interventions with low interindividual variation are likely to be widely generalisable, whilst an intervention with high interindividual variation is likely to have effects that are either subgroup or individual specific. The former kind of intervention might be widely applicable across individuals, whilst the latter kind of intervention requires specific research, typically with large samples (Hecksteden et al., 2015), to tease out subgroup- or participant-by-intervention interactions to facilitate successful practical application.

Comparison of heterogeneity in responses, such as post-scores or change scores to interventions, are not the only possible use of statistical methods for comparing variances. For example, in other fields such as ecology there have been calls to shift focus of analysis onto the exploration of dispersion of traits between groups in non-experimental or intervention designs (Nakagawa & Schielzeth, 2012). Some recent examples from sport and exercise science, and RT in particular, include primary research exploring between-participant acute response variation for the purposes of identifying methods⁹ to reduce RT stimulus heterogeneity, (Exner et al., 2022) as well as a meta-analysis exploring between-participant heterogeneity of accuracy in predicting proximity to task failure during RT (Halperin et al., 2022) and in the number of repetitions that can be performed at different percentages of one repetition maximum (Nuzzo et al., 2023b).

Given the value of embracing and exploring variation alongside mean effects in sport and exercise science, yet the lack of application in research synthesis by way of primary research or meta-analysis, we present and discuss effect size approaches and models for meta-analysis of variation. Indeed, meta-analysis presents a very valuable method for exploring variation in a field such as sport and exercise science due to the typically small samples in primary studies. Such small samples have even lower statistical power to detect differences in variation as compared to means (Yang et al., 2022).

Although ultimately we will recommend the thoughtful consideration of assumptions for certain approaches regarding what effect size statistics and models to employ in examining variation through use of meta-analysis, we provide examples throughout using all approaches described in order to aid the reader in understanding their strengths and weaknesses. We hope this will make clear why we offer such recommendations in our discussion and conclusion. As we have up to this point, we will also make an effort to both present the mathematical formulations of the models described, as well as to provide an explanation of them in plain English.

2 Effect size statistics for meta-analytic comparisons of variation

Until recent years there has been a dearth of effect size statistics available for the examination of variation in a meta-analytic framework. However, several have been proposed that we now describe: the standard deviation for individual responses (SD_{ir} ; Hopkins (2015); Atkinson & Batterham (2015); Atkinson et al. (2019)), the log ratio of standard deviations ($\ln VR$; termed the “variability ratio”; Hedges & Nowell (1995)), and the log ratio of coefficient of variation ($\ln CVR$; termed the “coefficient of variation ratio”; Nakagawa et al. (2015); Senior et al. (2020)). We present the independent groups versions due to use of randomised controlled trials in our examples below, but note that dependent versions (i.e., for comparing related samples) also exist for $\ln VR$ and $\ln CVR$ (Senior et al., 2020).

2.1 Standard deviation for individual responses (SD_{ir})

In the context of *precision* or *personalised* approaches to exercise prescription the SD_{ir} has been proposed as an approach to determine the extent to which individual responses manifest by comparison of variation between two groups; control and intervention (Atkinson et al., 2019; Atkinson & Batterham, 2015; Hopkins, 2015). The standard deviation of change scores (post-intervention scores minus pre-intervention scores) within the intervention group reflects the gross combination of a number of sources of variation including: participant-by-intervention interactions (i.e., actual individual responsiveness or ‘trainability’), within-participant variability

⁹Exploration of methodological approaches and their impact on heterogeneity have also been explored in preclinical research (Usui et al., 2021).

in intervention response (i.e., variability in response to the same intervention administered to the same participant), and random error (i.e., from pre and post measurements; Hecksteden et al. (2015)). The standard deviation of change scores from the control group (assuming it is a non-intervention control group and not something like a ‘usual-care’ group) by contrast is assumed to only reflect random error¹⁰ (Hecksteden et al., 2015). As such, the difference in these standard deviations can be used to determine the extent to which additional variation has been introduced by the intervention and that might reflect individual responses. Whilst the SD_{ir} has been proposed and used primarily in the context of individual response variation to interventions, it should be noted that this kind of absolute comparison of variance between groups or conditions is not limited to such applications.

The SD_{ir} , and its sampling variance, $s_{SD_{ir}}^2$ are given by:

$$SD_{ir} = \sqrt{s_E^2 - s_C^2} \quad (12)$$

$$s_{SD_{ir}}^2 = 2 \left(\frac{s_E^4}{n_E - 1} + \frac{s_C^4}{n_C - 1} \right) \quad (13)$$

Thus, the SD_{ir} reflects a comparison of the absolute variance in change scores between control and intervention groups. However, a potential concern with the SD_{ir} is its potential to violate assumptions of normality, which is not the case for other effect size statistics such as $\ln VR$ and $\ln CVR$.

2.2 Log ratio of standard deviations ($\ln VR$)

A similar effect size statistic for the comparison of absolute variance between groups, and one that has had wide applications in more than just intervention response variability within fields such as ecology and evolution, is the $\ln VR$ (Hedges & Nowell, 1995; Nakagawa et al., 2015; Senior et al., 2020). An unbiased estimator of the natural logarithm of a population standard deviation ($\ln \sigma$), and its sampling variance, $s_{\ln \sigma}^2$ is given by:

$$\ln \hat{\sigma} = \ln s + \frac{1}{2(n-1)} \quad (14)$$

$$s_{\ln \hat{\sigma}}^2 = \frac{1}{2(n-1)} \quad (15)$$

where $\ln \hat{\sigma}$ is an estimate of $\ln \sigma$, and it is assumed with sufficiently large sample size and value of σ that $\ln \sigma$ is normally distributed with variance $s_{\ln \sigma}^2$. Given equations (14) and (15), the logarithm of the ratio of standard deviations of two groups, such as a control and intervention, the $\ln VR$, and its sampling variance, $s_{\ln VR}^2$ is given by:

$$\ln VR = \ln \left(\frac{s_E}{s_C} \right) + \frac{1}{2(n_E - 1)} - \frac{1}{2(n_C - 1)} \quad (16)$$

$$s_{\ln VR}^2 = \frac{1}{2} \left(\frac{n_C}{(n_C - 1)^2} + \frac{n_E}{(n_E - 1)^2} \right) \quad (17)$$

However, due to both SD_{ir} and $\ln VR$ being comparisons of absolute variance, they may find limited applicability where the mean of one group is larger than the comparison group (e.g., when \bar{x}_E is larger than \bar{x}_C). In this case, it is likely that the standard deviation will be larger in the group with the larger mean (e.g.,

¹⁰Though notably, in the case of health behaviour studies it may be the case that if someone volunteers for a study it could conceivably motivate them to alter various habits even when they are assigned to a control group thus influencing change scores.

s_E is larger than s_C). This mean-variance relationship is common for many variables and datasets¹¹ and to highlight this we provide examples below. They also assume constant measurement error over the range of values for the mean, which can impact their utility for examining response variation (Tenan et al., 2020).

2.3 Log ratio of coefficient of variation (lnCVR)

The coefficient of variation is the ratio of the standard deviation to the mean; therefore, comparison of the coefficient of variation between groups will identify whether standard deviations differ more, or less, than would be predicted by their difference in means where a mean-variance relationship is present. In essence, the coefficient of variation is a means of standardising the standard deviation against the mean such that the relative variation in an effect is expressed. The natural logarithm of the ratio between the coefficients of variation from two groups, the lnCVR is thus a more generally applicable effect size statistic for examining variability between groups. Considering equations (10) and (16), the lnCVR is given by:

$$\ln CVR = \ln\left(\frac{CV_E}{CV_C}\right) + \frac{1}{2(n_E - 1)} - \frac{1}{2(n_C - 1)} \quad (18)$$

where CV_E and CV_C are s_E/\bar{x}_E and s_C/\bar{x}_C respectively. Senior et al. (2020) derived the sampling variance, $s_{\ln CVR}^2$, as:

$$s_{\ln CVR}^2 = \frac{s_C^2}{n_C \bar{x}_C^2} + \frac{s_C^4}{2n_C^2 \bar{x}_C^4} + \frac{n_C}{(n_C - 1)^2} + \frac{s_E^2}{n_E \bar{x}_E^2} + \frac{s_E^4}{2n_E^2 \bar{x}_E^4} + \frac{n_E}{(n_E - 1)^2} \quad (19)$$

3 Examples using resistance training studies

As noted, to facilitate understanding for those new to examination of variation, we provide primary examples of the approaches presented using data from RT studies included in a recent meta-analysis published in the *Journal of Sport Sciences* (Polito et al., 2021). Here we have used their list of included studies and re-extracted data from 111 of these¹². All analysis examples were performed in R (version 4.2.1, ‘‘Funny-Looking Kid’’, The R Foundation for Statistical Computing, 2022) using the **metafor** package (Viechtbauer, 2010). The extracted dataset, analysis scripts, models, data summaries, and supplementary materials are available on the Open Science Framework (<https://osf.io/2h9ma/>) or the GitHub repository (<https://github.com/jamessteeleii/Meta-Analysis-of-Variation-in-Resistance-Training.git>).

Polito et al. (2021) conducted a systematic review and meta-analysis of randomised trials that included a RT intervention group(s) and a non-training control comparison group. Their analysis focused upon the SMD between the RT intervention group(s) and the control group from the studies included, with both overall effect estimate and moderator analyses (i.e., meta-regressions¹³) performed. Given that Polito et al. (2021) included only studies with a non-training control group, their study selection offers a unique context to examine variation of interindividual responses specifically by comparing the variances in change scores between the RT intervention groups(s) and control group. Table 1 shows the total sample size, along with the

¹¹For one clear example, see figure 1A in Vigostky et al. (2020) who show that the mean and standard deviation for baseline strength values typically scale with one another across most studies.

¹²The authors of the meta-analysis did not make their extracted data openly available, nor did they respond to our request for the extracted data. Further, their original analysis included 119 studies however we were unable to extract data for our analyses from 8 of these for a variety of reasons (e.g., only percentage change data was reported, no standard deviations for control groups reported).

¹³Regression analyses are likely familiar to most readers where in the simplest form they try to predict the value of some dependent variable from some independent variable(s). This can be extended to meta-analytic synthesis where the independent variables reflect characteristics associated with the effects included. For example, they may reflect characteristics of the sample in the study for which the effect was extracted such as age or sex, or they might reflect characteristics of the intervention received such as the dose or frequency of exposure.

Table 1: Sample sizes for resistance training and non training control groups for dataset.

Arm	Sample Size
RT	
All	2683
Minumum RT	5
Median RT	12
Maximum RT	59
CON	
All CON	2349
Minumum CON	4
Median CON	10
Maximum CON	44

Note:

RT = resistance training

CON = non-training control

median and range by group, across the included studies. Indeed, this highlights the typically small samples used in sport and exercise science, and thus low power to detect difference in both means and variances in individual studies (Yang et al., 2022), emphasising the value of meta-analysis to explore variation. Table 2 shows the study and participant characteristics.

3.1 Detecting the presence of interindividual response variation to resistance training intervention with absolute variance statistics

First we conducted a traditional SMD and $\ln RR$ based effect size¹⁴ meta-analysis to explore the effects of RT interventions compared to controls for strength and hypertrophy (i.e., muscle mass/size) outcomes¹⁵. Polito et al. (2021) originally used a normal random-effects meta-analysis as described in the introduction. However, the data we extracted were hierarchical in nature. Thus, as opposed to there being only the assumption that studies are a random effect, due to there being multiple outcomes measured within each arm in the studies (i.e., intervention group(s) and control group, within each study), and that in some studies there were multiple interventions examined, there is the additional assumption that must be included that both the intervention groups and effects also come from overarching distributions. Thus a multilevel mixed-effects meta-analysis model (Van den Noortgate et al., 2013) with cluster-robust variance estimation (Hedges et al., 2010) was used with random intercepts for study, arm¹⁶, and effect. This model then includes additional τ^2 terms for each of the levels and assigns weights appropriately given this and the clustering of effects within arms within studies. We can therefore describe the overall model as:

¹⁴It is worth noting that in the sport and exercise sciences, similarly to other fields that examine the effects of experimental intervention, the most common study design for testing or estimating intervention effects is the randomised pretest-posttest-control design (i.e., an intervention and control, or other intervention, group randomly allocated and measured pre- and post-exposure). We presented the SMD and $\ln RR$ effect sizes in equations (6) and (10) merely for simplicity in the introduction, but note that extension of these for such 2x2 (i.e., condition x time) study designs have been presented in detail elsewhere (see: Gurevitch et al., (2000); Morris et al., (2007); Morris (2008); Lajeunesse (2011, 2015)) and these are the effect sizes used in the meta-analyses referred to here.

¹⁵We also explored for signs of small study bias, including publication bias favouring the finding of intervention effects, for the SMDs given that the relative lack of awareness for variance based effect sizes in the field implies that they might have more influence over such biases. There did not appear to be any obvious small study bias in the dataset (see <https://osf.io/stqr3>).

¹⁶We use the term *arm* to refer to an intervention group-control group contrast to accommodate studies including multiple intervention groups. This is so as to not confuse the reader with the use of *group* to designate either the RT intervention group(s) or control group separately. Thus, in the instances of models using effect sizes relating to comparisons between an intervention group and control group (i.e., SMD, $\ln RR$, SD_{ir} , $\ln VR$, and $\ln CVR$) we calculate comparisons *between* each intervention group (i.e., arm) and the control group. Thus, where a study had for example two RT interventions and a control, two separate arms would be coded (RT intervention 1 compared to control, and RT intervention 2 compared to control). Data was coded such that study and arm had explicit nesting.

Table 2: Summary of study and participant characteristics.

Characteristic	Summary
TESTEX	7 (6, 8)
Age	33 (23, 66)
Proportion Male	100 (0, 100)
Weight	74 (68, 78)
BMI	26.62 (24.27, 27.34)
Training Status	
Trained	9 (4.6%)
Untrained	187 (95%)
Sample Type	
Clinical	5 (2.6%)
Healthy	191 (97%)
RT Intervention Only?	
N	9 (4.6%)
Y	187 (95%)
Duration (weeks)	12 (8, 16)
Weekly Frequency	3.00 (2.00, 3.00)
Number of Exercises	6 (2, 8)
Sets per Exercise	3.00 (2.50, 3.00)
Number of Repetitions	10.0 (8.0, 11.2)
Load (%1RM)	74 (65, 80)
Task Failure?	
N	29 (23%)
Y	95 (77%)

Note:

RT = resistance training;

Continuous variables are median (IQR);

Categorical variables are count (%);

Not all studies reported full descriptive data (see dataset; <https://osf.io/kg2z4>)

$$\hat{\theta}_{ijk} = (\theta + \tau_{(1)i} + \tau_{(2)j} + \tau_{(3)k}) + m_{ijk} \quad (20)$$

where $\hat{\theta}_{ijk}$ is the k th effect size, here the SMD or $\ln RR$, from the j th arm ($j = 1, 2, \dots, N_j$; where N_j is the number of arms) in the i th study ($i = 1, 2, \dots, N_i$; where N_i is the number of studies), θ is the intercept or overall mean of the effects, τ_i is the deviation from θ for the i th study, τ_j is the deviation for the j th arm, τ_k is the deviation for the k th effect, and m_{ijk} is the sampling error for each effect size normally distributed with $\sigma_{\hat{\theta}_{ijk}}^2$. This model is referred to as ‘mixed’ effects because of the presence of both fixed (θ), and random ($\tau_{(1)i}, \tau_{(2)j}, \tau_{(3)k}$) effects¹⁷. The main term in the model we are interested in is θ which is our estimate for the overall weighted average effect (i.e., $\hat{\theta}$).

We then fit the same model for the SD_{ir} and $\ln VR$ effect sizes for change scores (i.e., post-intervention minus pre-intervention scores) in order to explore how absolute variance in responses differed between RT interventions and controls. A positive SMD or $\ln RR$ would indicate that RT interventions produced greater improvements in outcomes compared to controls, whilst a positive SD_{ir} and $\ln VR$ would indicate that the introduction of the RT intervention increased variation in responses (i.e., change scores) compared to controls (i.e., suggests the presence of interindividual response variation).

The pattern of results from our models examining SMDs (figure 2) were similar to those reported by Polito et al. (2021), albeit with slightly lower estimates for both outcome types; possibly due to our use of a multilevel mixed-effects meta-analysis model that allowed for each individual effect size to be more appropriately weighted (the relative amount of heterogeneity between effects for each level is presented as the I^2 statistic). As might be expected, in comparison to non-training controls the RT interventions produced increases in both strength (SMD = 0.87 [95%CI: 0.77 to 0.97]; $I_{study}^2 = 57.32\%$, $I_{arm}^2 = 3\%$, $I_{effect}^2 = 11.95\%$) and hypertrophy outcomes (SMD = 0.34 [95%CI: 0.29 to 0.39]; $I_{study}^2 = 54.62\%$, $I_{arm}^2 = 0.62\%$, $I_{effect}^2 = 2.79\%$). Confidence intervals on the overall effects were precise for both outcomes though prediction intervals, indicating the range over which we might expect future estimates of effects to fall based on this evidence, for SMD estimates (see figure 2) were fairly wide and relative heterogeneity was fairly high mostly as a result of between-study variance (i.e., effects were more similar *within* studies and arms than *between* them).

For the $\ln RR$ results we exponentiated them and converted to percentages to be more interpretable. These were similar, with greater proportional increases in strength compared with hypertrophy (figure 3). Increases were seen for both strength ($\exp RR = 21.97$ [95%CI: 19.43 to 24.57]; $I_{study}^2 = 33.46\%$, $I_{arm}^2 = 0\%$, $I_{effect}^2 = 0\%$) and hypertrophy ($\exp RR = 5.39$ [95%CI: 4.44 to 6.35]; $I_{study}^2 = 12.97\%$, $I_{arm}^2 = 0\%$, $I_{effect}^2 = 0\%$). Confidence intervals were again precise for both outcomes, and whilst relative heterogeneity was lower compared to SMD models prediction intervals were still quite wide.

In addition to the SMD and $\ln RR$ results, both the SD_{ir} (figure 4) and $\ln VR$ (figure 5) were also positive for both strength ($SD_{ir} = 0.91$ [95%CI: 0.36 to 1.47]; $I_{study}^2 = 53.85\%$, $I_{arm}^2 = 0.04\%$, $I_{effect}^2 = 0\%$; $\ln VR = 0.9$ [95%CI: 0.77 to 1.02]; $I_{study}^2 = 56.36\%$, $I_{arm}^2 = 0.77\%$, $I_{effect}^2 = 27.36\%$) and hypertrophy outcomes ($SD_{ir} = 0.42$ [95%CI: 0.33 to 0.5]; $I_{study}^2 = 0.01\%$, $I_{arm}^2 = 40.15\%$, $I_{effect}^2 = 39.73\%$; $\ln VR = 0.5$ [95%CI: 0.4 to 0.6]; $I_{study}^2 = 41.21\%$, $I_{arm}^2 = 3.31\%$, $I_{effect}^2 = 33.59\%$) indicating that exposure to the RT interventions may have introduced additional variance over and above random error, potentially suggesting the presence of interindividual response variation. Although, heterogeneity across the models and levels (study, arm, effect) were again relatively large and quite varied.

This additional variance might support previous perspectives (Carpinelli, 2017) that the considerable variation in responses to RT interventions typically observed are due to ‘true’ interindividual response variation over and above the random error that occurs from pre- and post-intervention measurements (i.e., the variation is *detectable* independent of the random error). However, as noted, both the SD_{ir} and $\ln VR$ assume constant variance over values of the mean (i.e., that the variance is similar whether mean values are low or high). As we have seen from the SMD and $\ln RR$ models, RT interventions increase mean scores. Thus, if there is a mean-variance relationship in the data, an increase in the mean alone may be fully responsible for any

¹⁷Technically then the *random* effects model presented earlier is also a *mixed* effects model. It is traditionally referred to as the random-effects model though.

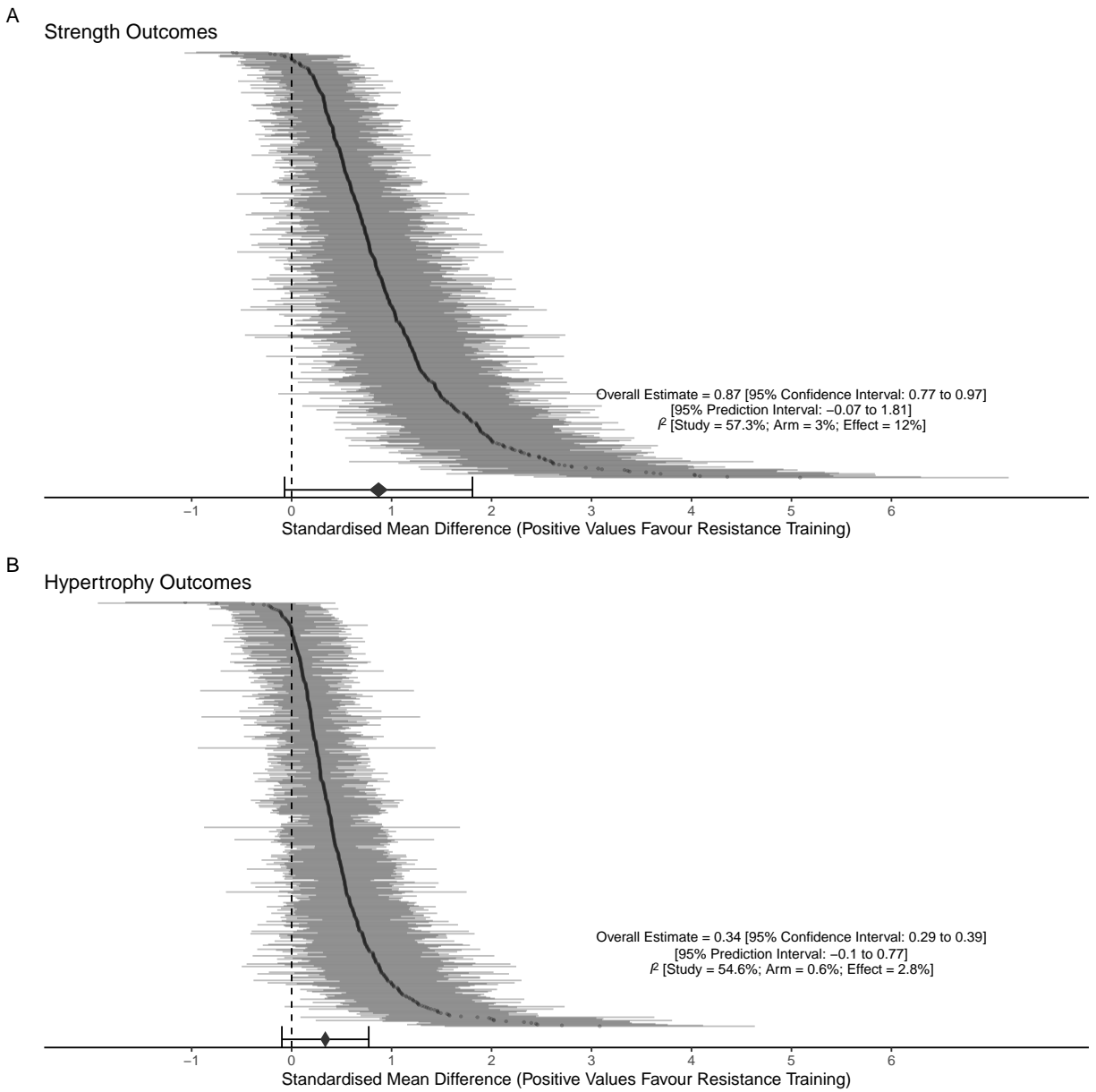


Figure 2: Caterpillar plots of SMD effect sizes for strength (A) and hypertrophy (B) outcomes.

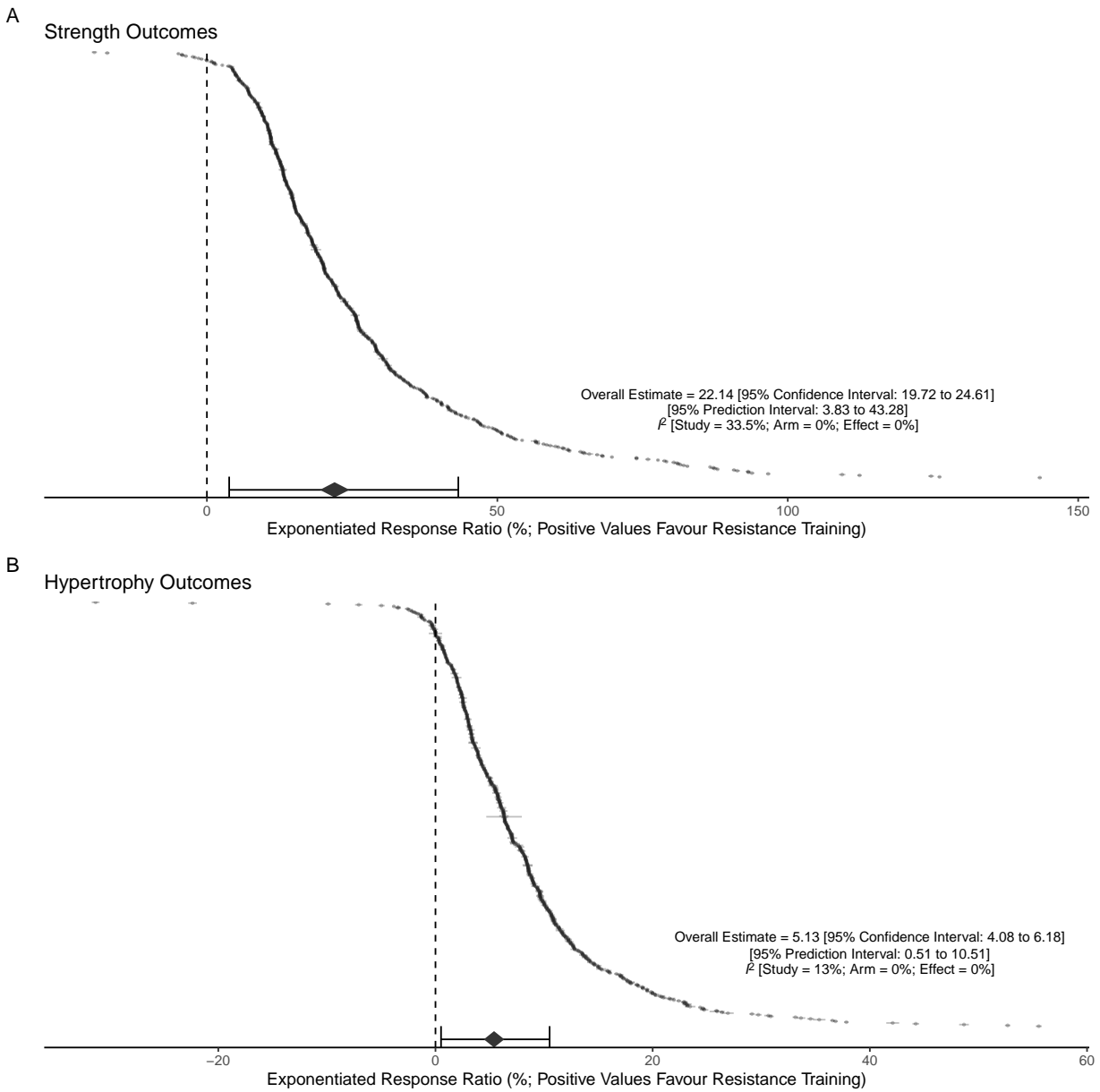


Figure 3: Caterpillar plots of exponentiated RR effect sizes for strength (A) and hypertrophy (B) outcomes.

apparent increase in variation. As such, we cannot rely solely on absolute comparisons of variance such as the SD_{ir} and $\ln VR$ to determine whether interindividual response variation is actually present. The $\ln CVR$ can be used to overcome this issue, and below we re-analyse this dataset using this effect size statistic. First though, we present data demonstrating the ubiquity of the mean-variance relationship in typical RT study outcome measures to emphasise the need to consider this assumption, and introduce a modelling approach that can also be used to overcome some possible limitations with the $\ln CVR$ and provide flexibility to accommodate wider applications.

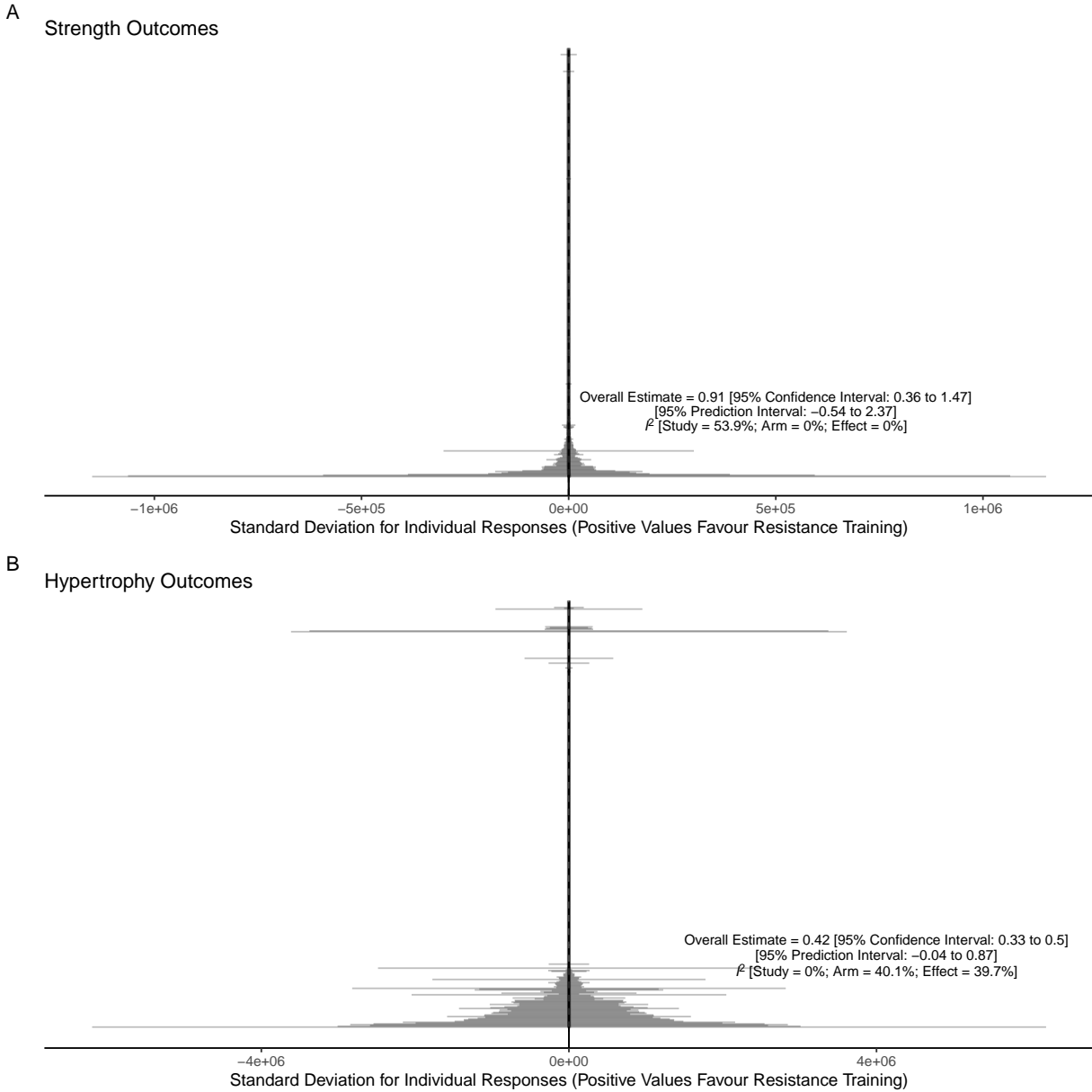


Figure 4: Caterpillar plots of SD_{ir} effect sizes for strength (A) and hypertrophy (B) outcomes.

3.2 Mean-variance relationships in muscular strength and hypertrophy

With meta-analytic models of variation we are not limited to solely exploring variation in responses to interventions (e.g., Halperin et al. (2022); Nuzzo et al. (2023b)). We can explore the relationships between

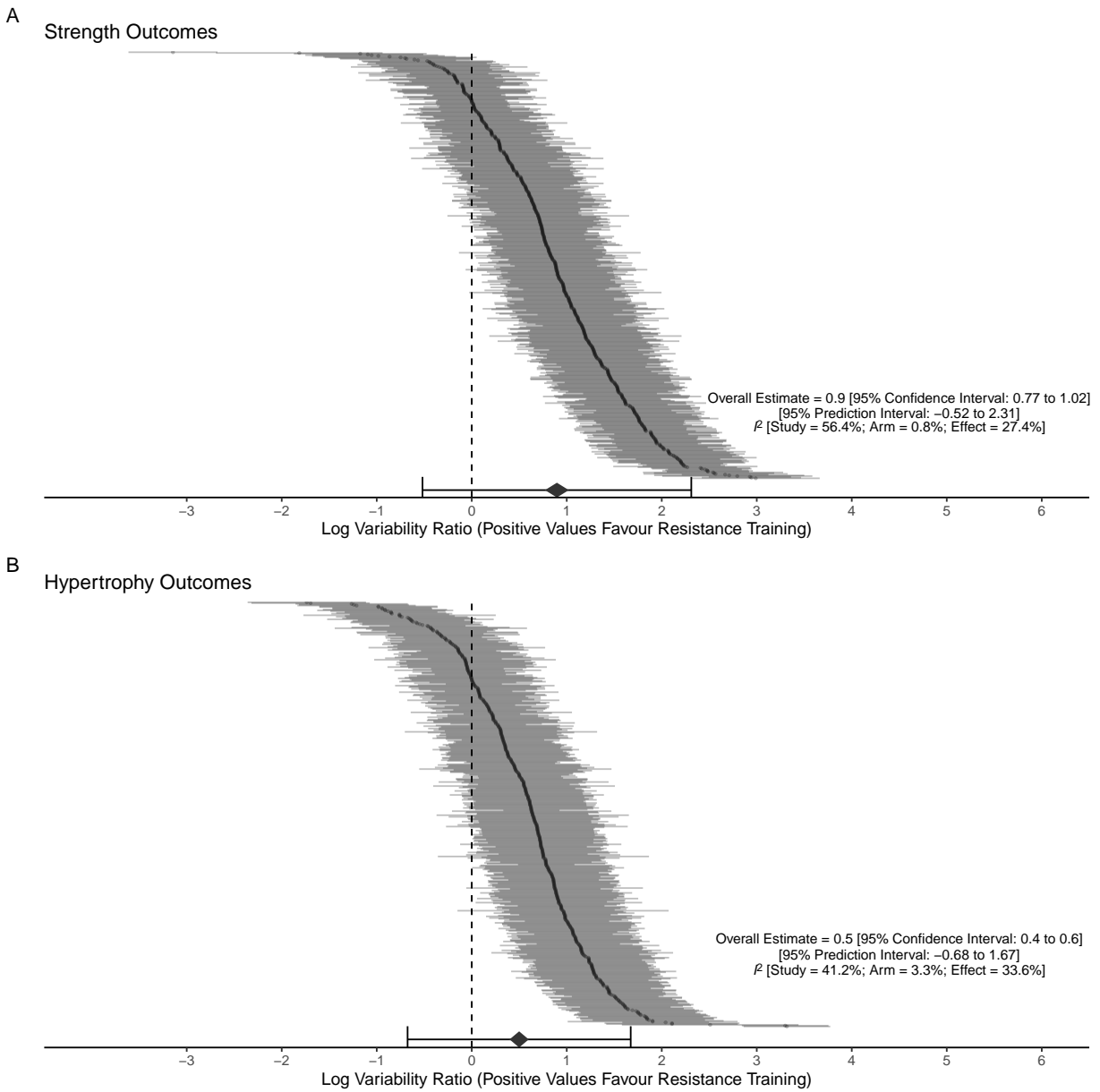


Figure 5: Caterpillar plots of $\ln VR$ effect sizes for strength (A) and hypertrophy (B) outcomes.

variance in a number of outcomes and the impact of certain predictors on this in the form of meta-regression. For example, as noted, one possible predictor of variance is the mean itself. As such, we can model variance of each effect as the response itself with the mean of the effect as the predictor. The standard deviation is, however, bounded at zero and so in many cases it may not conform to assumptions of normality which are required for regression models (i.e., that the residuals, the difference between the estimated and actual data, are normally distributed). Therefore, we instead can use $\ln \hat{\sigma}$, which is unbounded. In the following example we explore the mean-variance relationship in the pre-intervention scores for outcomes in the data set from Polito et al. (2021).

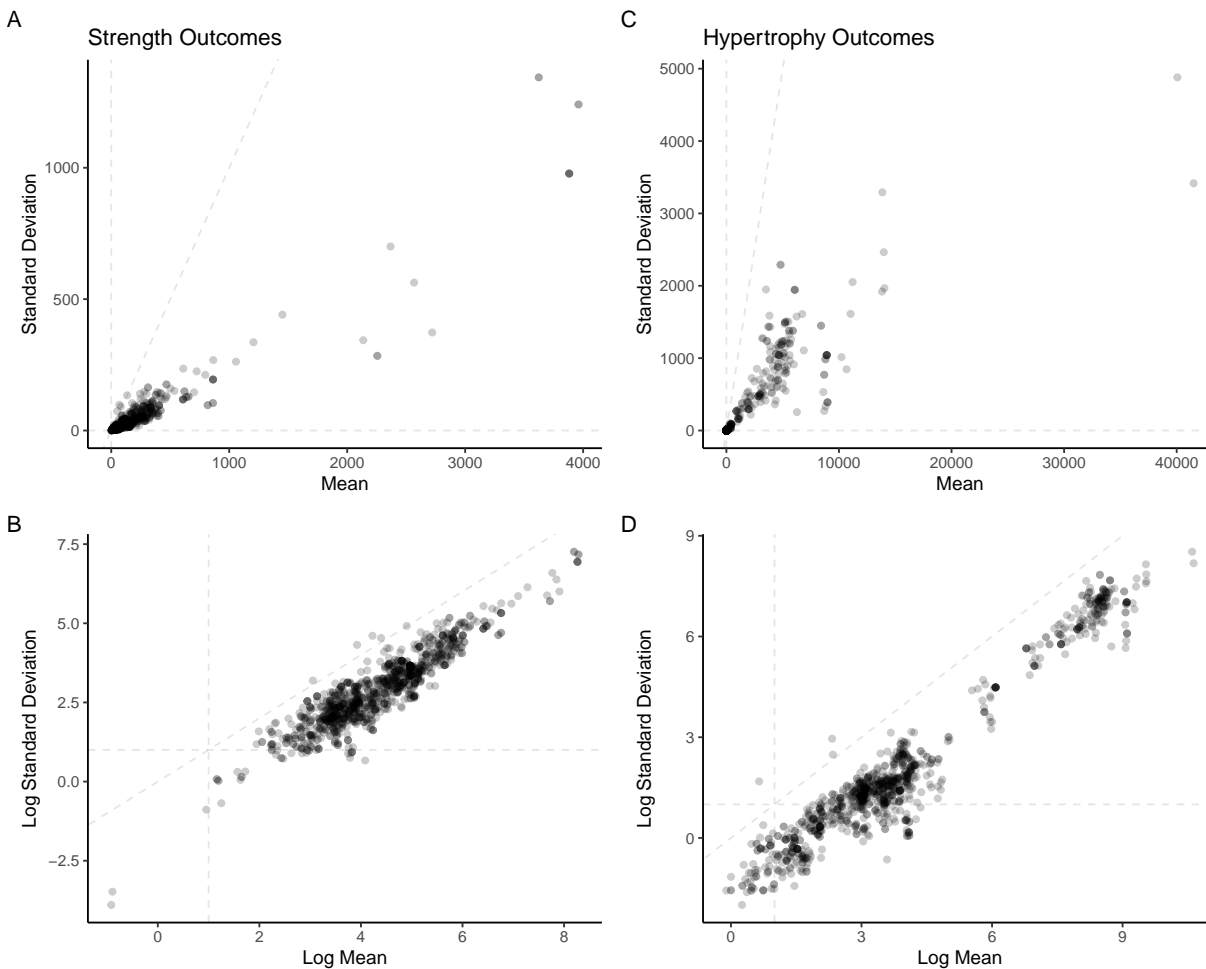


Figure 6: Scatter plots of raw mean and standard deviation of pre-intervention scores for (A) strength outcomes and (B) hypertrophy outcomes, and of the log mean and log standard deviation of pre-intervention scores for (C) strength outcomes and (D) hypertrophy outcomes.

As can be seen in figure 6(A) and (C), there is considerable heteroskedasticity in the relationship between the raw mean (\bar{x}) and standard deviation (s). The variance in standard deviations increases with higher mean values. This is similar to what is known as Taylor’s law in ecology, or the power law; in essence, an empirically derived relationship stating that the variance is a power function of the mean in many biological and physical systems (Taylor, 1961).

$$s^2 = a\bar{x}^b \quad (21)$$

where a and b are some constants. When this relationship holds, under most circumstances the standard

deviation is not proportional to the mean. However, when the mean and standard deviation are transformed to the log scale this relationship becomes linear based upon the product and power logarithmic rules:

$$2\ln s = \ln a + b\ln \bar{x} \quad (22)$$

Figure 6(B) and (D) shows that the relationship between the mean and variance on the log scale better meets the assumption of normality. Given these the observations we have for $\ln \hat{\sigma}$ and $\ln \bar{x}$ come from multiple outcomes within multiple arms within studies we can also estimate this relationship using a multilevel mixed-effects meta-regression model similar to that applied above. In this case though we are including an additional *predictor* variable, the $\ln \bar{x}$. For example, the following model specifies $\ln \bar{x}$ as a fixed effect with random intercepts for study, arm, and effect:

$$\ln \hat{\sigma}_{ijk} = (\beta_0 + \tau_{(1)i} + \tau_{(2)j} + \tau_{(3)k}) + \beta_1 \ln \bar{x}_{ijk} + \epsilon_{ijk} + m_{ijk} \quad (23)$$

where $\ln \hat{\sigma}_{ijk}$ is the k th effect size, as in equation (14), from the j th arm ($j = 1, 2, \dots, N_j$; where N_j is the number of arms¹⁸) in the i th study ($i = 1, 2, \dots, N_i$; where N_i is the number of studies), $\ln \bar{x}_{ijk}$ is the mean estimate for each effect size, β_0 is the intercept or overall mean of the effects, β_1 is the slope or regression coefficient for $\ln \bar{x}$, τ_i is the deviation from β_0 for the i th study, τ_j is the deviation for the j th arm, τ_k is the deviation for the k th effect, ϵ_{ijk} is the residual for each effect size which is normally distributed with σ_ϵ^2 , and m_{ijk} is the sampling error for each effect size normally distributed with $\sigma_{\ln \hat{\sigma}_{ijk}}^2$.

These kinds of models are highly flexible. Additional predictor terms could be added; for example, we could model a categorical variable for the outcome type and include $\beta_2 Outcome$ in the model with *Outcome* as a dummy coded variable for the outcome type (i.e., hypertrophy = 0, and strength = 1), where β_2 is the slope or regression coefficient for *Outcome* (most intuitively thought of as the difference between the two outcome types)¹⁹.

Figure 7 shows this model fit visually where the size of the points reflects their weight in the model. Both strength and hypertrophy outcomes show strong linearity between the mean and standard deviation on the log scale, though there is a small difference in intercepts between the two outcome types suggesting a slight but systematically greater degree of variance in strength measures compared to hypertrophy for a given mean score.

The presence of Taylor's law type relationships should be examined in datasets prior to deciding on which variance effect size statistic should be employed. Returning to the context of interindividual response variation to interventions, the presence of a mean-variance relationship in the data would imply that we cannot rely on absolute comparisons of variance (i.e., SD_{ir} or $\ln VR$) to determine whether interindividual response variation is actually present. So we should also explore this assumption for the change-scores in the RT and control groups and determining the appropriate effect sizes to explore.

¹⁸In contrast to the models presented examining effect sizes relating to comparisons between and intervention group and control group, in the models examining $\ln \hat{\sigma}_{ijk}$ and $\ln \bar{x}_{ijk}$ as a predictor the term *arm* refers to both the intervention groups(s) and control group. Thus, where a study had for example two RT interventions and a control, three separate arms would be coded (RT intervention 1, RT intervention 2, and control). Data were again coded such that study and arm had explicit nesting.

¹⁹We do not have to limit ourselves to only fixed effect predictor terms as we have here. Indeed, for mixed effect models generally some argue that models should use a *maximal* random effects structure including both random intercepts and slopes (i.e., that the effect of the predictor term can vary within different levels of the model and is also assumed to come from an overarching distribution of slopes), and their correlations, to enhance generalisability of inferences (Barr et al., 2013). We could model a categorical variable for the outcome type and using random effects include $(\beta_2 + \varphi_i) Outcome$ or $(\beta_2 + \varphi_i + \varphi_j) Outcome$ in the model with *Outcome* as a dummy coded variable for the outcome type (i.e., hypertrophy = 0, and strength = 1), where β_2 is the overall average slope or regression coefficient for *Outcome*, and φ_i is the deviation (random slope) from β_2 for the i th study and φ_j is the deviation for the j th arm. Indeed, we fit additional models using $\ln \hat{\sigma}$ with $\ln \bar{x}$ and *Outcome* as a predictor with (1) random intercepts for study and arm only, (2) the inclusion of correlated random slopes for $\ln \bar{x}$ by study, and (3) the inclusion of correlated random slopes for $\ln \bar{x}$ by study and arm. The comparison of these models is included in the supplementary materials (<https://osf.io/4xrcg>). Note, the addition of both random slopes for study, and for arm, improved model fit significantly, though we limit presentation in the main text to the simpler model.

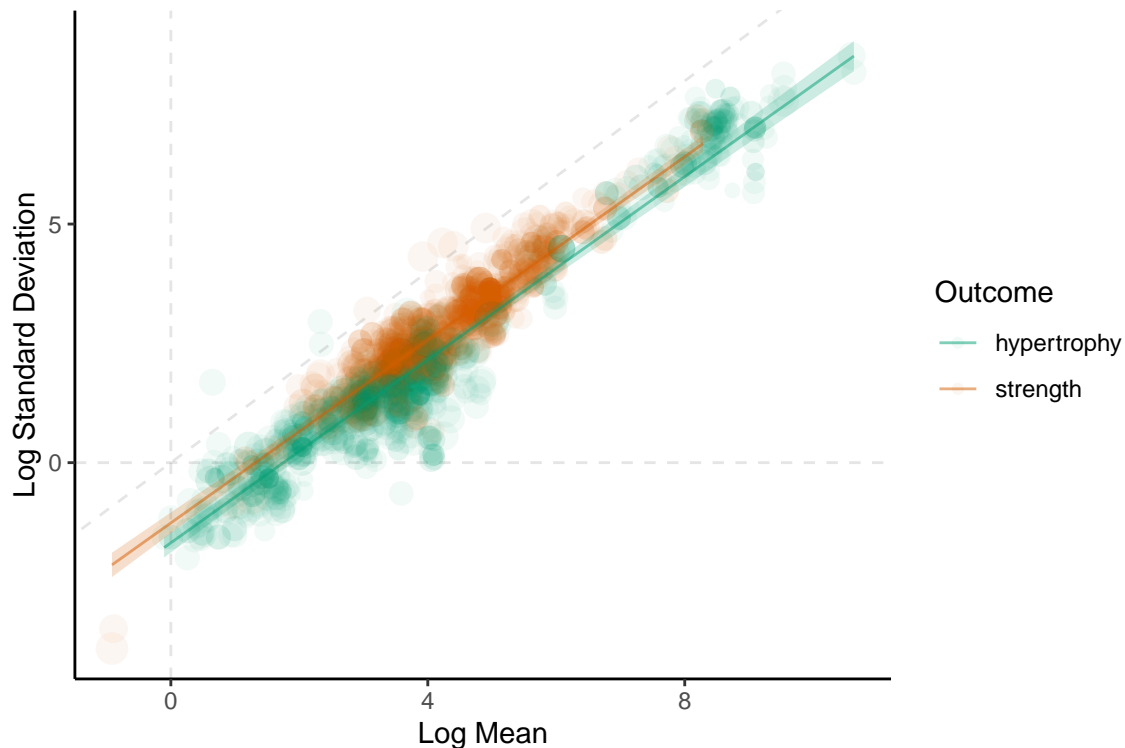


Figure 7: Meta-analytic scatter plot of the log mean and log standard deviation of pre-intervention scores.

3.3 Reanalysis of interindividual response variation using $\ln CVR$

As can be seen in figures 8(A) and (C) there is also a mean-variance relationship in the change score data about zero whereby an increase in the mean alone (i.e., greater mean change score in the intervention compared to the control) may be fully responsible for any apparent increase in variation. Further, when transforming change scores to absolute changes (i.e., converting all to positive numeric scores) we see that in figures 8(B) and (D) that the log transformation exhibits similar linearity as seen with the pre-intervention scores above. As such, in this case, we cannot rely solely on absolute comparisons of variance such as the SD_{ir} and $\ln VR$ to determine whether interindividual response variation is actually present.

The $\ln CVR$ can be used to overcome this issue though. Fitting the same multilevel mixed-effects meta-analysis model with cluster-robust variance estimation and random intercepts for study, arm, and effect as before (see equation (20)) using the $\ln CVR$ as the effect size statistic leads to different RT conclusions compared to absolute variance comparisons using SD_{ir} or $\ln VR$. The introduction of an RT intervention actually *reduces* the relative variation seen in change scores for strength ($\ln CVR = -0.61$ [95%CI: -0.76 to -0.47]; $I^2_{study} = 23.18\%$, $I^2_{arm} = 0\%$, $I^2_{effect} = 0\%$) and hypertrophy ($\ln CVR = -0.45$ [95%CI: -0.61 to -0.29]; $I^2_{study} = 10.03\%$, $I^2_{arm} = 0\%$, $I^2_{effect} = 0\%$) and further there is lower relative heterogeneity between studies in the effect estimates (figure 9).

There is, however, a potential limitation for the $\ln CVR$ also that may need to be considered. Firstly, it is limited to the use of ratio scale data (which is not the case for the $\ln \hat{\sigma}$ or $\ln VR$); hence the need to transform the change scores to be positively signed in this specific case. Secondly, whilst the $\ln CVR$ is useful in situations where there is a mean-variance relationship, the use of the CV in the effect size statistic assumes proportionality between standard deviation and mean. Where we see the kind of heteroskedasticity in the relationship between mean and standard deviation as we do for the change scores here (figure 8) an alternative yet comparable approach might be desirable. Lastly, this statistic is limited to examination of pairwise comparisons of variance. In this example this approach poses no issue as we are comparing RT intervention group(s) to a control group. But, as seen with the pre-score example above, alternative modelling

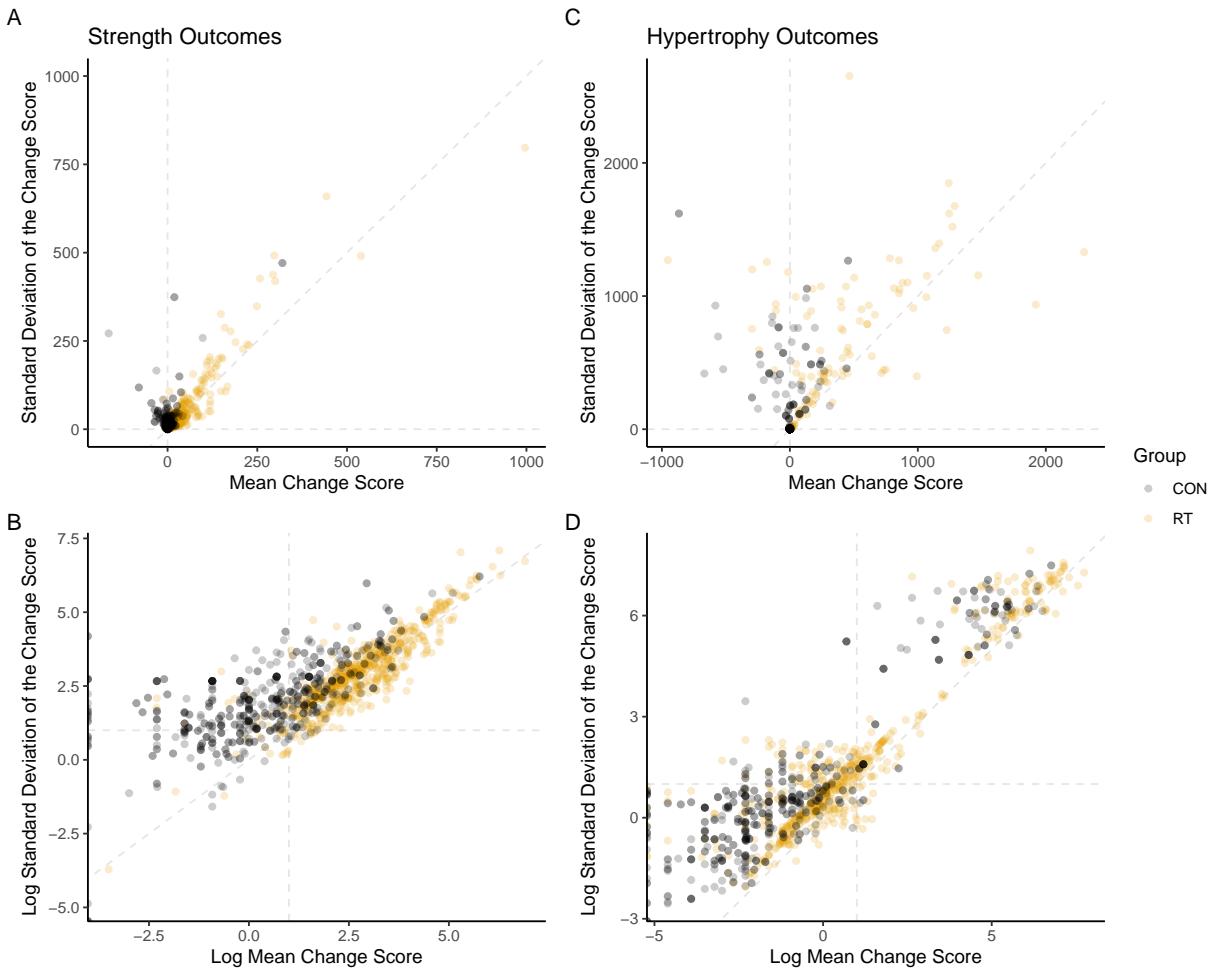


Figure 8: Scatter plots of raw mean and standard deviation of change scores for (A) strength outcomes and (B) hypertrophy outcomes, and of the log mean and log standard deviation of change scores for (C) strength outcomes and (D) hypertrophy outcomes.

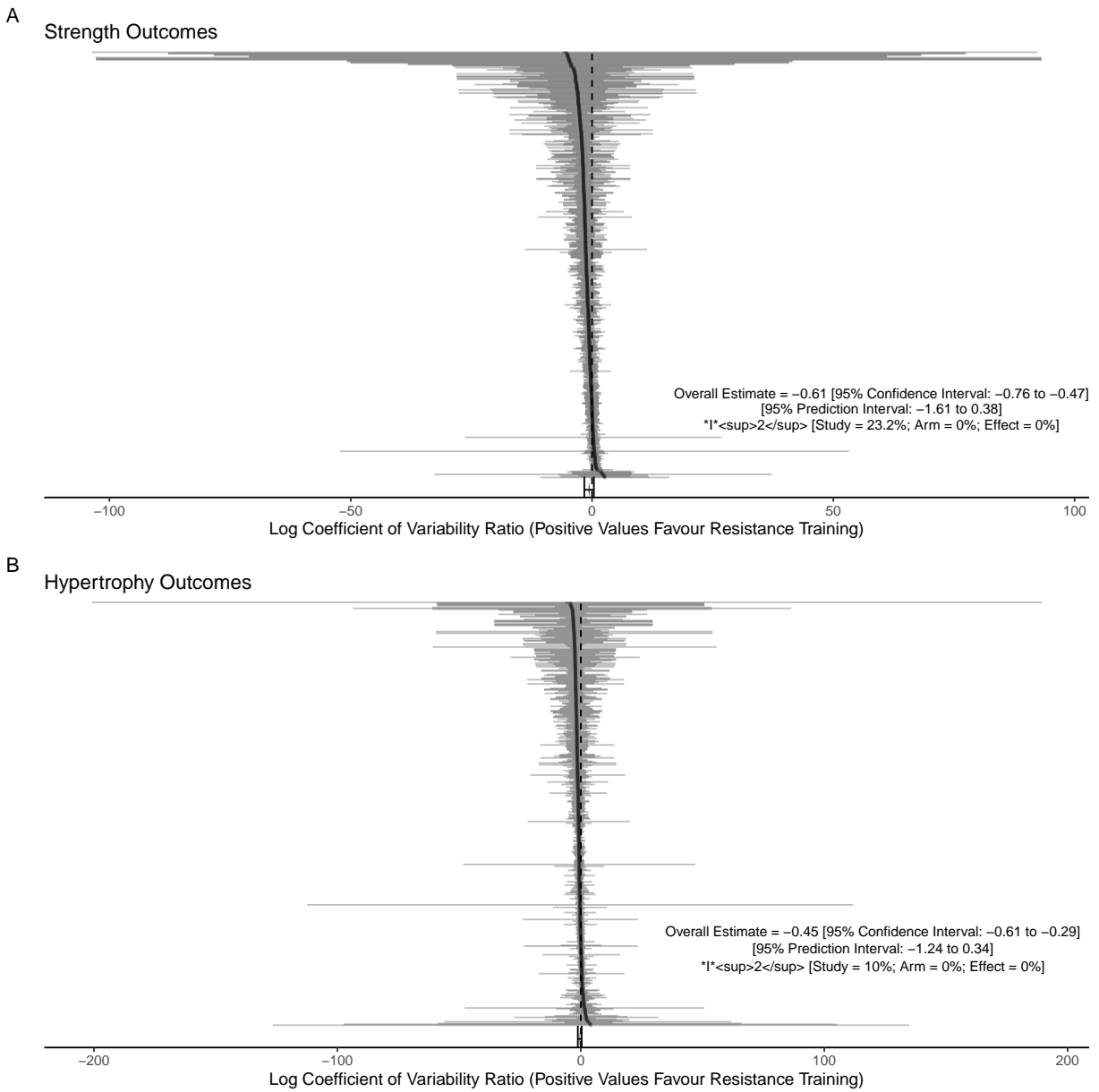


Figure 9: Caterpillar plots of $\ln CVR$ effect sizes for strength (A) and hypertrophy (B) outcomes.

approaches provide greater flexibility and can still offer estimation of a comparable effect of interest to the $\ln CVR$ one just applied.

3.4 Meta-regression of $\ln \hat{\sigma}$ with $\ln \bar{x}$ and $Group$

Instead, we can use a version of the meta-regression model described above (see equation (23) and the paragraph which followed it) to compare the variability in change scores between intervention and control groups using $\ln \hat{\sigma}$ and $\ln \bar{x}$. In this case, the categorical variable for the outcome type used previously is instead swapped for the group type and the new model term included becomes $\beta_2 Group$ with $Group$ as a dummy coded variable for the group (i.e., non-training control = 0, and RT intervention = 1), where β_2 is the slope or regression coefficient for $Group$. This model with just $Group$ as a predictor is comparable to the $\ln CVR$ model in the previous section where the $\beta_2 Group$ is the slope or regression coefficient for $Group$ and reflects the difference i.e., variance in the RT groups vs the control groups. This reflects the pairwise nature of the $\ln CVR$. Given the heteroskedasticity in the change scores means and standard deviations (see figure 8), we fit this model and other extensions²⁰ to the dataset (see figure 10 for the best fitting model plots).

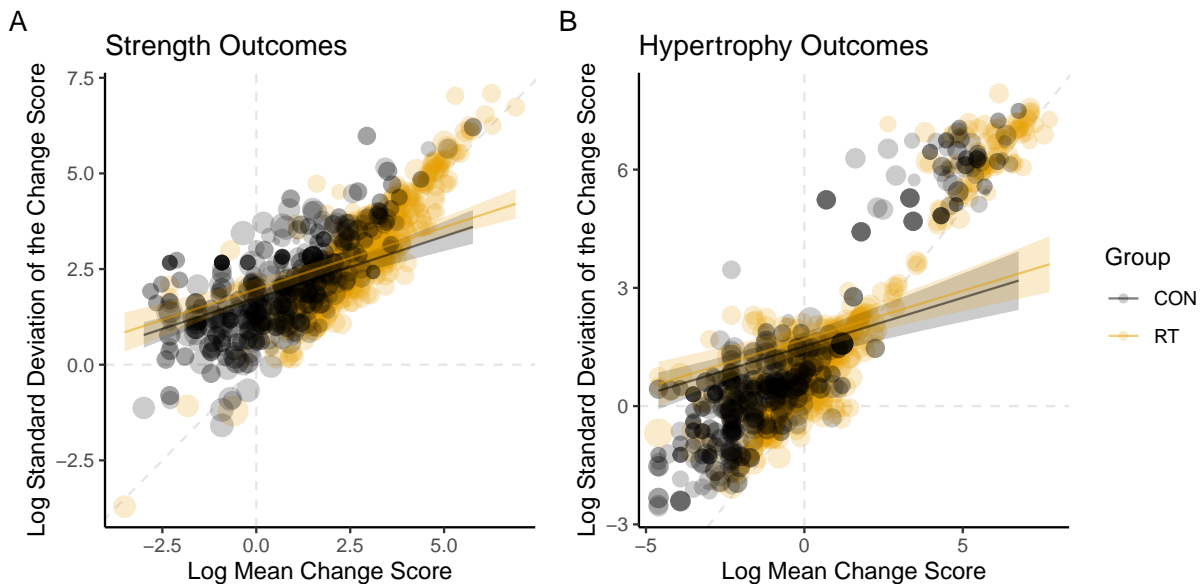


Figure 10: Meta-analytic scatter plot of the log mean and log standard deviation of change scores.

Despite the models comparability in terms of what they are trying to estimate (i.e., the difference in variance in change scores between the RT and CON groups) we arrive at substantively different conclusions compared to the $\ln CVR$ model and dependent upon the exact specification (see footnote 20) of the meta-regression model. Table 3 shows the different estimates from each model for this difference. Where the $\ln CVR$ suggested that the introduction of an RT intervention actually reduced variance in change scores (i.e., the estimate was negative) we see the opposite conclusion (i.e., the estimate is positive) for all multilevel meta-regression models of strength, and for hypertrophy this appears to be dependent on whether the coefficients (i.e., slopes) for the relationship between $\ln \hat{\sigma}$ and $\ln \bar{x}$ are allowed to vary between studies and arms.

It is worth explaining why this might be the case to aid readers in considering the assumptions that different models make. The $\ln VR$ and $\ln CVR$ models can be thought of as similar in that they both make fixed assumptions about the relationship between mean and variance; In the $\ln VR$ it is assumed to be zero, and

²⁰Note, as with the models examining *Outcome* upon baseline scores, we similarly explored $\ln \hat{\sigma}$ with $\ln \bar{x}$ and $Group$ as a predictor with (1) random intercepts for study and arm only, (2) the inclusion of random slopes for $\ln \bar{x}$ by study, and (3) the inclusion of random slopes for $\ln \bar{x}$ by study and arm. The comparison of these models is included in the supplementary materials (strength - <https://osf.io/n4wgk>; hypertrophy - <https://osf.io/hf8dx>). In this case, for strength the addition of random slopes for study, but not for arm, improved model fit significantly, and for hypertrophy intercept only model fit best.

Table 3: Comparison of estimates from model using $\ln CVR$ and multilevel meta-regression models of $\ln \hat{\sigma}$ and $\ln \bar{x}$ with group (RT vs CON) as a predictor

Outcome	Model	Estimate	Lower 95% CI	Upper 95% CI
Strength				
	$\ln CVR$	-0.61	-0.76	-0.47
	MLMR: Random Intercept	0.29	0.09	0.49
	MLMR: + Random Slope (Study)	0.24	0.04	0.43
	MLMR: + Random Slope (Study & Arm)	0.24	0.04	0.43
Hypertrophy				
	$\ln CVR$	-0.45	-0.61	-0.29
	MLMR: Random Intercept	0.18	0.01	0.34
	MLMR: + Random Slope (Study)	-0.34	-0.64	-0.03
	MLMR: + Random Slope (Study & Arm)	-0.34	-0.65	-0.04

Note:

CI = confidence interval

MLMR = multilevel meta-regression

in the $\ln CVR$ it is assumed to be proportional i.e., one. In both however this is a strong assumption. The multilevel meta-regressions on the other hand actually *estimate* this relationship (i.e., the value of β_1 , the slope or regression coefficient for $\ln \bar{x}$) and in the random slope models this is also estimated allowing it to vary between studies and arms (i.e., in some studies there may be a more or less strong relationship compared to others). Evidently, the assumptions that are made about the data-generating processes can impact upon the modelling strategy chosen²¹. Mean-variance relationships are important to consider when exploring variance effects, but it is also important to consider whether or not this relationship is assumed to be some fixed proportional value (i.e., as the $\ln CVR$ does) or whether or not this should be estimated from the data and whether it might also vary across studies and arms (i.e., as the multilevel meta-regression model allows). It is not necessarily clear which model should be preferred here, but thought should be given to the assumptions each makes and the fit of each model to the data.

3.5 Moderators of variance effects

Hopefully it is clear from the meta-regression models here, where we have included both fixed and random predictors as both categorical (i.e., *Outcome*, or *Group*) and continuous (i.e., $\ln \bar{x}$) variables, that there is considerable flexibility in the inclusion of predictors when exploring variance through a meta-analytic framework. Of course, the pairwise models described can also be extended to meta-regressions to explore not only how study, arm, or effect level characteristics moderate effect size estimates when considering effect sizes such as SMDs or $\ln RR$, but also when considering the variance-based effect size statistics and models employed in this article (i.e., SD_{ir} , $\ln VR$, $\ln CVR$, and $\ln \hat{\sigma}$)²². But these are limited to the pairwise comparisons of a categorical variable as the effect size. The meta-regression models presented here for $\ln \hat{\sigma}$ allow for comparisons to be extended beyond two categories including any number of arbitrary predictors, fixed and random and assumptions about their correlations.

²¹It should also be noted that these models all assume that the \bar{x} is estimated without error which is clearly not the case. Given that for most effects that might be included in such models we can determine the sampling variance for \bar{x} one approach to address this might be to employ models that incorporate the variance on this predictor (i.e., measurement error or errors in variables models). This is beyond the scope of this paper to discuss.

²²See supplementary materials (<https://osf.io/e6vpr>) for examples from model estimates for both SMD and $\ln CVR$, (used for simplicity of presenting moderator analysis results) across a range of categorical and continuous predictors for both strength and hypertrophy outcomes. There were no obvious moderators of $\ln CVR$ in particular.

4 Discussion

Given the apparent lack of awareness of the utility of meta-analytic frameworks for exploring variance, and the potential value such analyses can offer for the sport and exercise sciences, we have presented some existing effect size statistics and models pertinent to this topic that hopefully will encourage and support researchers in the field to embrace more than just the mean when engaging in quantitative evidence synthesis. Indeed, for a field such as sport and exercise science where sample sizes are typically small, meta-analysis becomes even more valuable as such small samples in primary studies have even lower statistical power to detect differences in variation as compared to means²³ (Yang et al., 2022).

It is of particular interest to note the different conclusions drawn here dependent on the approach taken to determine from non-training control and RT intervention data whether or not there is *detectable* inter-individual response variation present. We deliberately presented these varying approaches to highlight their assumptions and to aid in readers understanding of their applications. Given the different conclusions drawn from the examples provided, we recommend that researchers consider whether assumptions of simpler approaches are met, or indeed seem reasonable, before their application in specific situations. Where for example there is not an obvious mean-variance relationship it may be appropriate to utilise the SD_{ir} or $\ln VR$. However, when this is present in the data then the $\ln CVR$ or meta-regression of $\ln \hat{\sigma}$ upon $\ln \bar{x}$ may be more appropriate, though assumptions regarding whether this relationship is proportional and fixed should also be considered. The structure of the data also impacts the specific modelling approach to be employed and whether or not assumptions that the mean-variance relationship varies within study or arm levels can be incorporated. Further, depending on the exact research questions it is worth considering balancing model complexity with its ability to provide an answer. If a simple pairwise comparison across a categorical variable is of interest then this can be explored with comparable models using the $\ln CVR$ or meta-regression of $\ln \hat{\sigma}$ upon $\ln \bar{x}$. But where more complicated predictors, including categories extending beyond just two, are of interest then the flexibility of the latter is desirable to explore. Lastly, given the possible sensitivity of substantive conclusions to the exact model specifications used it is worth considering the use of so called *multiverse* approaches to exploring questions such as those pertaining to variation with meta-analysis. With multiverse approaches, instead of having to choose one modelling approach, researchers use the many possible approaches (such as presented here) and explore how sensitive their substantive conclusions might be to nuances of model specifications and assumptions (Olsson-Collentine et al., 2023; Steegen et al., 2016).

Whether or not there is clear evidence for average intervention effects, research aimed at exploring variation is important. Even in the absence of an average intervention effect it may be that a large enough variance could imply that the intervention is at least beneficial to some (Usui et al., 2021). Such results might lead researchers to consider that further research exploring subgroup- or participant-by-intervention interactions is required to maximise successful practical application of such an intervention to avoid negative effects for some, and ensure positive effects for others. Interventions, which induce both meaningful average treatment effects and also show little evidence suggestive of interindividual variation, are likely to be widely generalisable and so from a practical perspective might offer considerable value in that we can have high expectations that everyone receiving them will likely improve (Usui et al., 2021); that is to say we can assume a constant effect and that the average intervention effect is indicative of the individual intervention effect (Cortés Martínez, 2021). Interventions such as these are valuable for the simplification of guidelines and recommendations. This also potentially represents another interesting area of future study regarding variation opposite to the usual search for individual variation; specifically, how to produce interventions that actually reduce variance in an outcome. In other contexts such as sporting performance, interventions to not only positively affect mean performance but also those that reduce variation in performance would be highly desirable.

The examples presented herein used data from RT studies included in a recent meta-analysis published in the *Journal of Sport Sciences* (Polito et al., 2021), which hopefully makes them more relatable for researchers in sport and exercise sciences. Given the varied results, this dataset does not necessarily provide clear answers as to the question of whether true interindividual response variation is present and so this remains an open area of research. However, the examples provided highlight the importance of considering carefully the

²³Indeed, it can be seen from figures 4, 5, and 9 that many of the individual study effect estimates have very large sampling errors.

assumptions inherent in the modelling approaches chosen to explore such a substantive topic. We hope they also prompt researchers to further explore variation.

5 Conclusion

Embracing variability and focusing on more than merely the mean differences between groups or conditions, such as intervention and control comparisons, has the potential to inform experimental design and lead to changes in both the approach and direction of follow-up studies. Whether there is evidence of meaningful average intervention effects or not, where considerable variance effects are present it suggests that a meaningful line of research would be to aim at identifying subgroup- or participant-by-intervention interactions using appropriate study designs (Hecksteden et al., 2015). Where variance effects are limited this instead suggests that translational work towards generalisable implementation might be the most meaningful line of future research. Finally, there may be cases where it is in fact desirable to identify interventions that actually reduce variance; for example, improvements in methodological approaches to enhance research (Usui et al., 2021), or interventions to reduce variation in sport performances. Thus, researchers in sport and exercise science should consider exploring variance more systematically, and indeed utilise the meta-analytic framework to support this. This could include the re-analysis of past meta-analyses as we have done here, and indeed researchers conducting future meta-analyses in the field of sport and exercise science should consider the value of concomitantly exploring means and variances utilising the established approaches (Atkinson et al., 2019; Atkinson & Batterham, 2015; Hopkins, 2015; Mills et al., 2021; Nakagawa et al., 2015; Usui et al., 2021) presented here and echoing the efforts of other recent work (Bonafiglia et al., 2022; Esteves et al., 2021; Fisher et al., 2022; Kelley et al., 2020, 2022; Steele et al., 2021). Researchers should be aware though that the meta-analytic modelling of variance, or indeed any effect size statistics, requires careful consideration of the research question and assumptions that different models make. Substantive conclusions may be sensitive to subtle differences in the approaches we have presented.

6 References

- Atkinson, G., & Batterham, A. M. (2015). True and false interindividual differences in the physiological response to an intervention. *Experimental Physiology*, *100*(6), 577–588. <https://doi.org/10.1113/EP085070>
- Atkinson, G., Williamson, P., & Batterham, A. M. (2019). Issues in the determination of 'responders' and 'non-responders' in physiological research. *Experimental Physiology*, *104*(8), 1215–1225. <https://doi.org/10.1113/EP087712>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 10.1016/j.jml.2012.11.001. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bernárdez-Vázquez, R., Raya-González, J., Castillo, D., & Beato, M. (2022). Resistance Training Variables for Optimization of Muscle Hypertrophy: An Umbrella Review. *Frontiers in Sports and Active Living*, *4*. <https://www.frontiersin.org/articles/10.3389/fspor.2022.949021>
- Bonafiglia, J. T., Swinton, P. A., Ross, R., Johannsen, N. M., Martin, C. K., Church, T. S., Slentz, C. A., Ross, L. M., Kraus, W. E., Walsh, J. J., Kenny, G. P., Goldfield, G. S., Prud'homme, D., Sigal, R. J., Earnest, C. P., & Gurd, B. J. (2022). Interindividual Differences in Trainability and Moderators of Cardiorespiratory Fitness, Waist Circumference, and Body Mass Responses: A Large-Scale Individual Participant Data Meta-analysis. *Sports Medicine (Auckland, N.Z.)*. <https://doi.org/10.1007/s40279-022-01725-9>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. <https://doi.org/10.1002/jrsm.12>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. John Wiley & Sons.
- Caldwell, A., & Vigotsky, A. D. (2020). A case against default effect sizes in sport and exercise science. *PeerJ*, *8*, e10314. <https://doi.org/10.7717/peerj.10314>
- Carpinelli, R. N. (2017). Interindividual heterogeneity of adaptations to resistance training. *Medicina Sportiva Practica*, *18*(4), 79–94.

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Cortés Martínez, J. (2021). Constant effect in randomized clinical trials with quantitative outcome : A methodological review. *TDX (Tesis Doctorals En Xarxa)*. <https://upcommons.upc.edu/handle/2117/349575>
- Curtis, P. S., & Wang, X. (1998). A meta-analysis of elevated CO₂ effects on woody plant mass, form, and physiology. *Oecologia*, *113*(3), 299–313. <https://doi.org/10.1007/s004420050381>
- Deb, S. K., Brown, D. R., Gough, L. A., Mclellan, C. P., Swinton, P. A., Andy Sparks, S., & Mcnaughton, L. R. (2018). Quantifying the effects of acute hypoxic exposure on exercise performance and capacity: A systematic review and meta-regression. *European Journal of Sport Science*, *18*(2), 243–256. <https://doi.org/10.1080/17461391.2017.1410233>
- Esteves, G. P., Swinton, P., Sale, C., James, R. M., Artioli, G. G., Roschel, H., Gualano, B., Saunders, B., & Dolan, E. (2021). Individual Participant Data Meta-Analysis Provides No Evidence of Intervention Response Variation in Individuals Supplementing With Beta-Alanine. *International Journal of Sport Nutrition and Exercise Metabolism*, *31*(4), 305–313. <https://doi.org/10.1123/ijsem.2021-0038>
- Exner, R. J., Patel, M. H., Whitener, D. V., Buckner, S. L., Jessee, M. B., & Dankel, S. J. (2022). Does performing resistance exercise to failure homogenize the training stimulus by accounting for differences in local muscular endurance? *European Journal of Sport Science*, 1–10. <https://doi.org/10.1080/17461391.2021.2023657>
- Fisher, J., Steele, J., Wolf, M., Korakakis, P. A., Smith, D., & Giessing, J. (2022). The Role of Supervision in Resistance Training; an Exploratory Systematic Review and Meta-Analysis: *International Journal of Strength and Conditioning*, *2*(1). <https://doi.org/10.47206/ijsc.v2i1.101>
- Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, *5*(10), 3–8. <https://doi.org/10.2307/1174772>
- Gurevitch, J., Morrison, J. A., & Hedges, L. V. (2000). The Interaction between Competition and Predation: A Meta-analysis of Field Experiments. *The American Naturalist*, *155*(4), 435–453. <https://doi.org/10.1086/303337>
- Hagger, M. (2006). Meta-analysis in sport and exercise research: Review, recent developments, and recommendations. *European Journal of Sport Science*, *6*(2), 103–115. <https://doi.org/10.1080/17461390500528527>
- Hagger, M. (2022). Meta-analysis. *International Review of Sport and Exercise Psychology*, *15*(1), 120–151. <https://doi.org/10.1080/1750984X.2021.1966824>
- Halperin, I., Malleron, T., Har-Nir, I., Androulakis-Korakakis, P., Wolf, M., Fisher, J., & Steele, J. (2022). Accuracy in Predicting Repetitions to Task Failure in Resistance Exercise: A Scoping Review and Exploratory Meta-analysis. *Sports Medicine*, *52*(2), 377–390. <https://doi.org/10.1007/s40279-021-01559-x>
- Hecksteden, A., Kraushaar, J., Scharhag-Rosenberger, F., Theisen, D., Senn, S., & Meyer, T. (2015). Individual response to exercise training - a statistical perspective. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *118*(12), 1450–1459. <https://doi.org/10.1152/jappphysiol.00714.2014>
- Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The Meta-Analysis of Response Ratios in Experimental Ecology. *Ecology*, *80*(4), 1150–1156. [https://doi.org/10.1890/0012-9658\(1999\)080%5B1150:TMAORR%5D2.0.CO;2](https://doi.org/10.1890/0012-9658(1999)080%5B1150:TMAORR%5D2.0.CO;2)
- Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science (New York, N. Y.)*, *269*(5220), 41–45. <https://doi.org/10.1126/science.7604277>
- Hedges, L. V., & Olkin, I. (2014). *Statistical Methods for Meta-Analysis*. Academic Press.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. <https://doi.org/10.1002/jrsm.5>
- Heidel, K. A., Novak, Z. J., & Dankel, S. J. (2022). Machines and free weight exercises: A systematic review and meta-analysis comparing changes in muscle size, strength, and power. *The Journal of Sports Medicine and Physical Fitness*, *62*(8), 1061–1070. <https://doi.org/10.23736/S0022-4707.21.12929-9>
- Hopkins, W. G. (2015). Individual responses made easy. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *118*(12), 1444–1446. <https://doi.org/10.1152/jappphysiol.00098.2015>
- Hrubeniuk, T. J., Bonafiglia, J. T., Bouchard, D. R., Gurd, B. J., & Sénéchal, M. (2022). Directions for Exercise Treatment Response Heterogeneity and Individual Response Research. *International Journal of Sports Medicine*, *43*(1), 11–22. <https://doi.org/10.1055/a-1548-7026>
- Kelley, G. A. (2022). Precision exercise medicine in rheumatology: Don't put the cart before the horse.

- Clinical Rheumatology*, 41(8), 2277–2279. <https://doi.org/10.1007/s10067-022-06260-6>
- Kelley, G. A., Kelley, K. S., & Callahan, L. F. (2022). Are There Interindividual Differences in Anxiety as a Result of Aerobic Exercise Training in Adults with Fibromyalgia? An Ancillary Meta-analysis of Randomized Controlled Trials. *Archives of Physical Medicine and Rehabilitation*, S0003-9993(22)00007-7. <https://doi.org/10.1016/j.apmr.2021.12.019>
- Kelley, G. A., Kelley, K. S., & Pate, R. R. (2020). Are There Inter-Individual Differences in Fat Mass and Percent Body Fat as a Result of Aerobic Exercise Training in Overweight and Obese Children and Adolescents? A Meta-Analytic Perspective. *Childhood Obesity (Print)*, 16(5), 301–306. <https://doi.org/10.1089/chi.2020.0056>
- Lajeunesse, M. J. (2011). On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology*, 92(11), 2049–2055. <https://doi.org/10.1890/11-0423.1>
- Lajeunesse, M. J. (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*, 96(8), 2056–2063. <https://doi.org/10.1890/14-2402.1>
- Mills, H. L., Higgins, J. P. T., Morris, R. W., Kessler, D., Heron, J., Wiles, N., Davey Smith, G., & Tilling, K. (2021). Detecting Heterogeneity of Intervention Effects Using Analysis and Meta-analysis of Differences in Variance Between Trial Arms. *Epidemiology (Cambridge, Mass.)*, 32(6), 846–854. <https://doi.org/10.1097/EDE.0000000000001401>
- Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, 11(2), 364–386. <https://doi.org/10.1177/1094428106291059>
- Morris, W. F., Hufbauer, R. A., Agrawal, A. A., Bever, J. D., Borowicz, V. A., Gilbert, G. S., Maron, J. L., Mitchell, C. E., Parker, I. M., Power, A. G., Torchin, M. E., & Vázquez, D. P. (2007). Direct and interactive effects of enemies and mutualists on plant performance: A meta-analysis. *Ecology*, 88(4), 1021–1029. <https://doi.org/10.1890/06-0442>
- Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society*, 82(4), 591–605. <https://doi.org/10.1111/j.1469-185X.2007.00027.x>
- Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, 6(2), 143–152. <https://doi.org/10.1111/2041-210X.12309>
- Nakagawa, S., & Schielzeth, H. (2012). The mean strikes back: Mean-variance relationships and heteroscedasticity. *Trends in Ecology & Evolution*, 27(9), 474–475; author reply 475–476. <https://doi.org/10.1016/j.tree.2012.04.003>
- Nuzzo, J., Pinto, M. D., Nosaka, K., & Steele, J. (2023a). The Eccentric:Concentric Strength Ratio of Human Skeletal Muscle In Vivo: Meta-analysis of the Influences of Sex, Age, Joint Action, and Velocity. *Sports Medicine*, 53(6), 1125–1136. <https://doi.org/10.1007/s40279-023-01851-y>
- Nuzzo, J., Pinto, M., Nosaka, K., & Steele, J. (2023b). *Maximal number of repetitions at percentages of the one repetition maximum: A meta-regression and moderator analysis of sex, age, training status, and exercise*. SportRxiv. <https://doi.org/10.51224/SRXIV.291>
- Olsson-Collentine, A., Van Aert, R. C. M., Bakker, M., & Wicherts, J. (2023). Meta-analyzing the multiverse: A peek under the hood of selective reporting. *Psychological Methods*. <https://doi.org/10.1037/met0000559>
- Pickering, C., & Kiely, J. (2019). Do Non-Responders to Exercise Exist-and If So, What Should We Do About Them? *Sports Medicine (Auckland, N.Z.)*, 49(1), 1–7. <https://doi.org/10.1007/s40279-018-01041-1>
- Plackett, R. L. (1958). Studies in the History of Probability and Statistics: VII. The Principle of the Arithmetic Mean. *Biometrika*, 45(1/2), 130–135. <https://doi.org/10.2307/2333051>
- Polito, M. D., Papst, R. R., & Farinatti, P. (2021). Moderators of strength gains and hypertrophy in resistance training: A systematic review and meta-analysis. *Journal of Sports Sciences*, 39(19), 2189–2198. <https://doi.org/10.1080/02640414.2021.1924978>
- Ross, R., Goodpaster, B. H., Koch, L. G., Sarzynski, M. A., Kohrt, W. M., Johannsen, N. M., Skinner, J. S., Castro, A., Irving, B. A., Noland, R. C., Sparks, L. M., Spielmann, G., Day, A. G., Pitsch, W., Hopkins, W. G., & Bouchard, C. (2019). Precision exercise medicine: Understanding exercise response variability. *British Journal of Sports Medicine*, 53(18), 1141–1153. <https://doi.org/10.1136/bjsports-2018-100328>
- Senior, A. M., Viechtbauer, W., & Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Research Synthesis Methods*, 11(4), 553–567. <https://doi.org/10.1002/jrsm.1423>

- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Steele, J., Plotkin, D., Van Every, D., Rosa, A., Zambrano, H., Mendelovits, B., Carrasquillo-Mercado, M., Grgic, J., & Schoenfeld, B. J. (2021). Slow and Steady, or Hard and Fast? A Systematic Review and Meta-Analysis of Studies Comparing Body Composition Changes between Interval Training and Moderate Intensity Continuous Training. *Sports (Basel, Switzerland)*, 9(11), 155. <https://doi.org/10.3390/sports9110155>
- Swinton, P. A., Hemingway, B. S., Saunders, B., Gualano, B., & Dolan, E. (2018). A Statistical Framework to Interpret Individual Response to Intervention: Paving the Way for Personalized Nutrition and Exercise Prescription. *Frontiers in Nutrition*, 5, 41. <https://doi.org/10.3389/fnut.2018.00041>
- Swinton, P. A., Katherine, B., Andy, H., Leon, G., John, P., Rodrigo, R. A., Patrick, M., & Andrew, M. (2022). Interpreting magnitude of change in strength and conditioning: Effect size selection, threshold values and Bayesian updating Journal of Sports Sciences. *Journal of Sports Sciences*.
- Taylor, L. R. (1961). Aggregation, Variance and the Mean. *Nature*, 189(4766), 732–735. <https://doi.org/10.1038/189732a0>
- Tenan, M. S., Vigotsky, A. D., & Caldwell, A. R. (2020). Comment on: “A Method to Stop Analyzing Random Error and Start Analyzing Differential Responders to Exercise.” *Sports Medicine*, 50(2), 431–434. <https://doi.org/10.1007/s40279-019-01249-9>
- Usui, T., Macleod, M. R., McCann, S. K., Senior, A. M., & Nakagawa, S. (2021). Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. *PLOS Biology*, 19(5), e3001009. <https://doi.org/10.1371/journal.pbio.3001009>
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, 45(2), 576–594. <https://doi.org/10.3758/s13428-012-0261-6>
- Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, 36, 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Vigotsky, A., Nuckols, G. L., Fisher, J., Heathers, J., Krieger, J., Schoenfeld, B. J., Giessing, J., & Steele, J. (2020). *Improbable data patterns in the work of Barbalho et al.* SportRxiv. <https://doi.org/10.31236/osf.io/sg3wm>
- Wolf, M., Androulakis-Korakakis, P., Fisher, J., Schoenfeld, B., & Steele, J. (2023). Partial Vs Full Range of Motion Resistance Training: A Systematic Review and Meta-Analysis. *International Journal of Strength and Conditioning*, 3(1). <https://doi.org/10.47206/ijsc.v3i1.182>
- Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I., & Nakagawa, S. (2022). Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Global Change Biology*, 28(3), 969–989. <https://doi.org/10.1111/gcb.15972>