Meta-Analysis of Variation in Sport and Exercise Science

Examples of Application Within Resistance Training Research

James Steele James Fisher Dave Smith Andrew Vigotsky
Brad Schoenfeld Yefeng Yang Shinichi Nakagawa

2022-08-04

Abstract

TO COMPLETE.

1 Introduction

Though the quantitative synthesis of results across studies has existed since the 17th century (Plackett 1958), the modern day term "meta-analysis" as it is now referred to was coined by Gene Glass in (1976). Since that time use of meta-analysis as a tool for the synthesis of research in sport and exercise science has increased considerably (Hagger 2022). One area of sport and exercise science which has comprised a reasonable proportion of that growth in recent years appears to have been in the study of resistance training (RT; figure 1). Given such growth, throughout the paper we use RT studies as a hopefully familiar example for sport and exercise science researchers.

As with many other fields (Nakagawa et al. 2015; Usui et al. 2021; Mills et al. 2021) likely the most familiar aim with the use of meta-analysis, and indeed primary empirical research too, in sport and exercise science is to make comparisons between the means of measurements taken across different categorical grouping variables; for example, the comparison of an intervention group and a control group, the comparison of one intervention group to 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.

Most commonly a magnitude based[1] effect size statistic (Caldwell and Vigotsky 2020), the standardised mean difference (SMD), is used to compare means between groups or conditions. This is usually Cohen's d (Cohen 1988), or it's bias-corrected metric referred to as Hedges' g (Larry V. Hedges and Olkin 2014; Borenstein et al. 2021; Nakagawa and Cuthill 2007) [2]. The SMD, and its sampling variance, s^2_{SMD} are given by:

$$SMD = \frac{\overline{x}_E - \overline{x}_C}{s_{pooled}} J \tag{1}$$

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

$$s_{pooled} = \sqrt{\frac{(n_C - 1)s_C^2 + (n_E - 1)s_E^2}{n_C + n_E - 2}} \tag{3}$$

[1]Though notably not all meta-analyses use *magnitude based* effect sizes. Indeed some explicitly use what Caldwell and Vigotsky term *signal-to-noise* effect sizes (e.g., Heidel, Novak, and Dankel (2022)).

[2]We will refer to both merely as the SMD throughout the manuscript for simplicity.

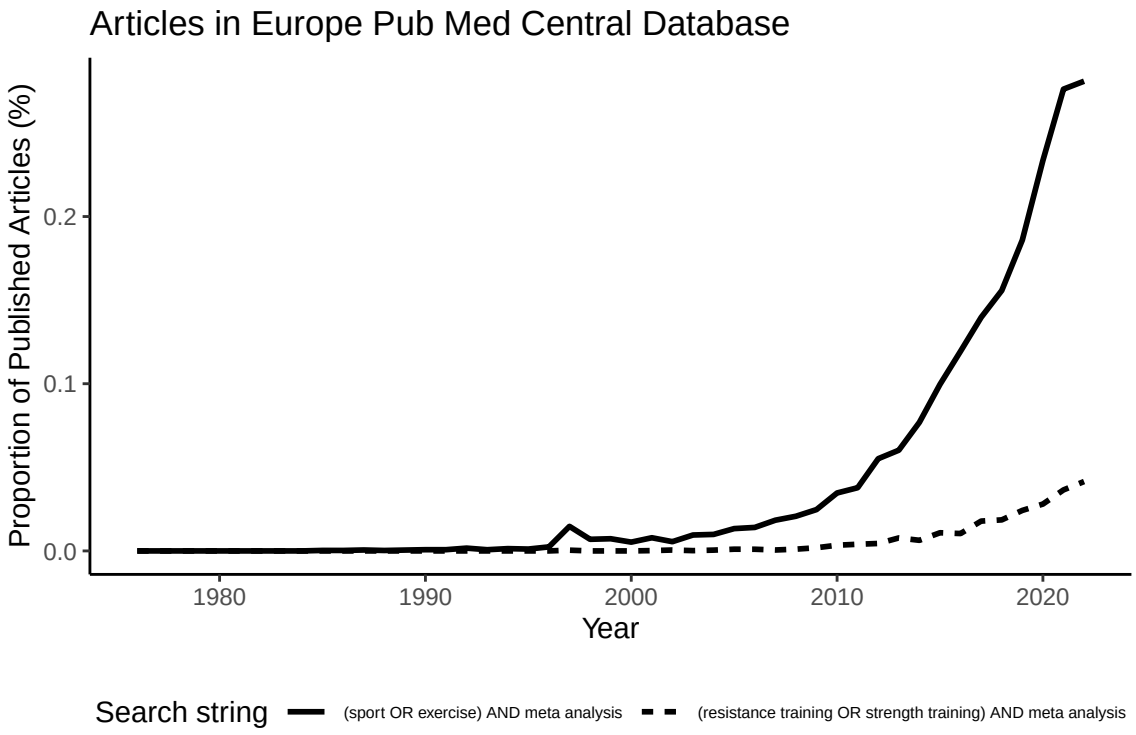

Figure 1: Trends in meta-analyses published in sport and exercise science since 1976.

$$s^2_{SMD} = \frac{n_C + n_E}{n_C n_E} + \frac{SMD^2}{2(n_E + n_C)} \tag{4}$$

where $\overline{x}_C$ and $\overline{x}_E$ 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 and Wang 1998; Larry V. Hedges, Gurevitch, and Curtis 1999; Lajeunesse 2015). The lnRR, and its sampling variance, $s^2_{\ln RR}$ are given by:

$$\ln RR = \ln \frac{\overline{x}_E}{\overline{x}_C} \tag{5}$$

$$s^2_{\ln RR} = \frac{s^2_C}{n_C \overline{x}^2_C} + \frac{s^2_E}{n_E \overline{x}^2_E} \tag{6}$$

Due to its calculation the SMD is affected not only by the difference in means of the two groups, but is also affected by the standard deviations of both groups due to the standardisation of the effect size by s_{pooled}. Contrastingly, the $\ln RR$ is uninfluenced by the standard deviations in either groups (see equation (5)), which only affects the sampling variance (see equation (6)). Despite this, use of $\ln RR$ has been limited in previous meta-analyses in sport and exercise science (Deb et al. 2018) and to our knowledge only one meta-analysis in RT has used this effect size (Swinton et al., in press).

Though the focus in sport and exercise science among other fields has been in estimating the average effects of interventions for both primary research and synthesis through meta-analysis, the field has been aware for some

time that responses to certain interventions do vary potentially on a subgroup or even individual basis. The increased interested 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 (Hecksteden et al. 2015; Atkinson and Batterham 2015; Atkinson, Williamson, and Batterham 2019; Ross et al. 2019; Swinton et al. 2018; Hopkins 2015; Kelley 2022; Hrubeniuk et al. 2022; Pickering and Kiely 2019). However, despite the availability of approaches to compare variances between groups, in sports and exercise science this is rarely explored in primary research (Bonafiglia et al. 2022) and, though there has been increased interest in recent years, few meta-analyses in sport and exercise include both comparisons of means and variances or explicitly aim to investigate the latter (Kelley, Kelley, and Callahan 2022; Kelley, Kelley, and Pate 2020; Esteves et al. 2021; Bonafiglia et al. 2022; Steele et al. 2021; Fisher et al. 2022). 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 to tease apart subgroup- or participant-by-intervention interactions to enable successful practical application.

Comparison of heterogeneity in responses such as post-scores or change scores to interventions are however 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 onto the exploration of dispersion of traits between groups in non-experimental or interventional designs (Nakagawa and 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 methods[3] to reduce RT stimulus heterogeneity (Exner et al. 2022) as well as meta-analysis exploring between participant heterogeneity of accuracy in predicting proximity to task failure during RT (Halperin et al. 2022).

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 meta-analysis, we present and discuss effect size approaches and models for meta-analysis of variation. We provide examples of the approaches presented using data from RT studies included in a recent meta-analysis published in the *Journal of Sport Sciences* (Polito, Papst, and Farinatti 2021).

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 and Batterham (2015); Atkinson, Williamson, and Batterham (2019)), the log ratio of standard deviations ($\ln VR$; termed the "variability ratio"; L. V. Hedges and Nowell (1995)), and the log ratio of coefficient of variation ($\ln CVR$; termed the "coefficient of variation ratio"; Nakagawa et al. (2015)).

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

In the context of *precision* or *personalised* approaches to exercise prescription comparison of variation between two groups, control (C) and intervention (E), the SD_{ir} has been proposed as an approach to determining the extent to which individual responses manifest (Hopkins 2015; Atkinson and Batterham 2015; Atkinson, Williamson, and Batterham 2019). The standard deviation of changes 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

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

'trainability'), within-participant variability 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 only reflects random error (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 that might reflect individual responses. The SD_{ir}, and its sampling variance, $s^2_{SD_{ir}}$ are given by:

$$SD_{ir} = \sqrt{s_E^2 - s_C^2} \tag{7}$$

$$s^2_{SD_{ir}} = 2(\frac{s_E^4}{n_E - 1} + \frac{s_C^4}{n_C - 1}) \tag{8}$$

Thus, the SD_{ir} reflects a comparison of the absolute variance in change scores between C and E. 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.

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

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

$$\ln\hat{\sigma} = \ln s + \frac{1}{2(n-1)} \tag{9}$$

$$s^2_{\ln\hat{\sigma}} = \frac{1}{2(n-1)} \tag{10}$$

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 $s^2_{\ln\sigma}$. Given equations (9) and (10), the logarithm of the ratio of standard deviations of two groups, such as a control (C) and intervention (E), the $\ln VR$, and its sampling variance, $s^2_{\ln VR}$ is given by:

$$\ln VR = \ln(\frac{s_E}{s_C}) + \frac{1}{2(n_E - 1)} - \frac{1}{2(n_C - 1)} \tag{11}$$

$$s^2_{\ln VR} = \frac{1}{2(n_E - 1)} + \frac{1}{2(n_C - 1)} \tag{12}$$

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 $\overline{x}_E$ is larger than $\overline{x}_C$). In this case it is likely that the standard deviation will be larger in the group with the larger mean (e.g., s_E is larger than s_C). This mean-variance relationship is common for many variables and datasets[4] and we will provide examples of this 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, Vigotsky, and Caldwell 2020).

[4]For 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.

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. 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 (5) and (11), the lnCVR is given by:

$$\ln CVR = \ln(\frac{CV_E}{CV_C}) + \frac{1}{2(n_E - 1)} - \frac{1}{2(n_C - 1)} \tag{13}$$

where CV_E and CV_C are $s_E/\overline{x}_E$ and $s_C/\overline{x}_C$ respectively. Nakagawa et al. (2015) derived the sampling variance, $s^2_{\ln CVR}$, as:

$$\begin{aligned} s^2_{\ln CVR} = &\frac{s^2_C}{n_C\overline{x}_C} + \frac{1}{2(n_C - 1)} - 2\rho_{\ln\overline{x}_C,\ln s_C}\sqrt{\frac{s^2_C}{n_C\overline{x}_C} + \frac{1}{2(n_C - 1)}} \\ &+\frac{s^2_E}{n_E\overline{x}_E} + \frac{1}{2(n_E - 1)} - 2\rho_{\ln\overline{x}_E,\ln s_E}\sqrt{\frac{s^2_E}{n_E\overline{x}_E} + \frac{1}{2(n_E - 1)}} \end{aligned} \tag{14}$$

where $\rho_{\ln\overline{x}_C,\ln s_C}$ and $\rho_{\ln\overline{x}_E,\ln s_E}$ are the correlatios between the means and the standard deviation in the C and E groups respectively on the log scale across studies.

3 Examples using resistance training studies

As noted, the examples presented used data from RT studies included in a recent meta-analysis published in the *Journal of Sport Sciences* (Polito, Papst, and Farinatti 2021). Here we have used their list of included studies and re-extracted data from 111 of these[5]. All analysis examples were performed in using 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 all available on the Open Science Framework (https://osf.io/2h9ma/).

Polito et al. (2021) conducted a systematic review and meta-analysis of randomised trials including a RT intervention where a non-training control comparison group was included. Their analysis focused upon the SMD between the RT interventions and control groups from the studies included, with both overall effect estimate and moderator analyses (i.e., meta-regressions) were performed. Given that Polito et al. (2021) included only studies with non-training control groups, their study selection offers a unqiue context to examine variation of interindividual responses specifically by means of comparing the variances in change scores between the RT intervention and control groups. Table 1 shows the total sample size, along with the median and range by group, across the included studies. Table 2 shows the study and participant characteristics.

3.1 Detecting the presence of interindividual response variation to resistance training intervention

First we conduct a traditional SMD based effect size meta-analysis to explore the effects of RT compared to control for strength outcomes and hypertrophy (i.e., muscle mass/size) outcomes[6]. Polito et al. (2021)

[5]The authors of the meta-analysis did not make their extracted data openly available, not 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).

[6]Note, we 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

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

Group	Sample Size
RT	
All RT	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

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 + Adjuvant Intervention?	
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)

originally used a normal random-effects meta-analysis, however the data we extracted were hierarchical in nature (multiple effects within groups within studies) and so a multilevel mixed-effects meta-analysis model with cluster-robust variance estimation was used with random intercepts for study and group[7]. 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 in responses differed between RT interventions and controls. A positive SMD would indicate that RT interventions increase 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 compared to controls (i.e., suggests the presence of interindividual response variation).

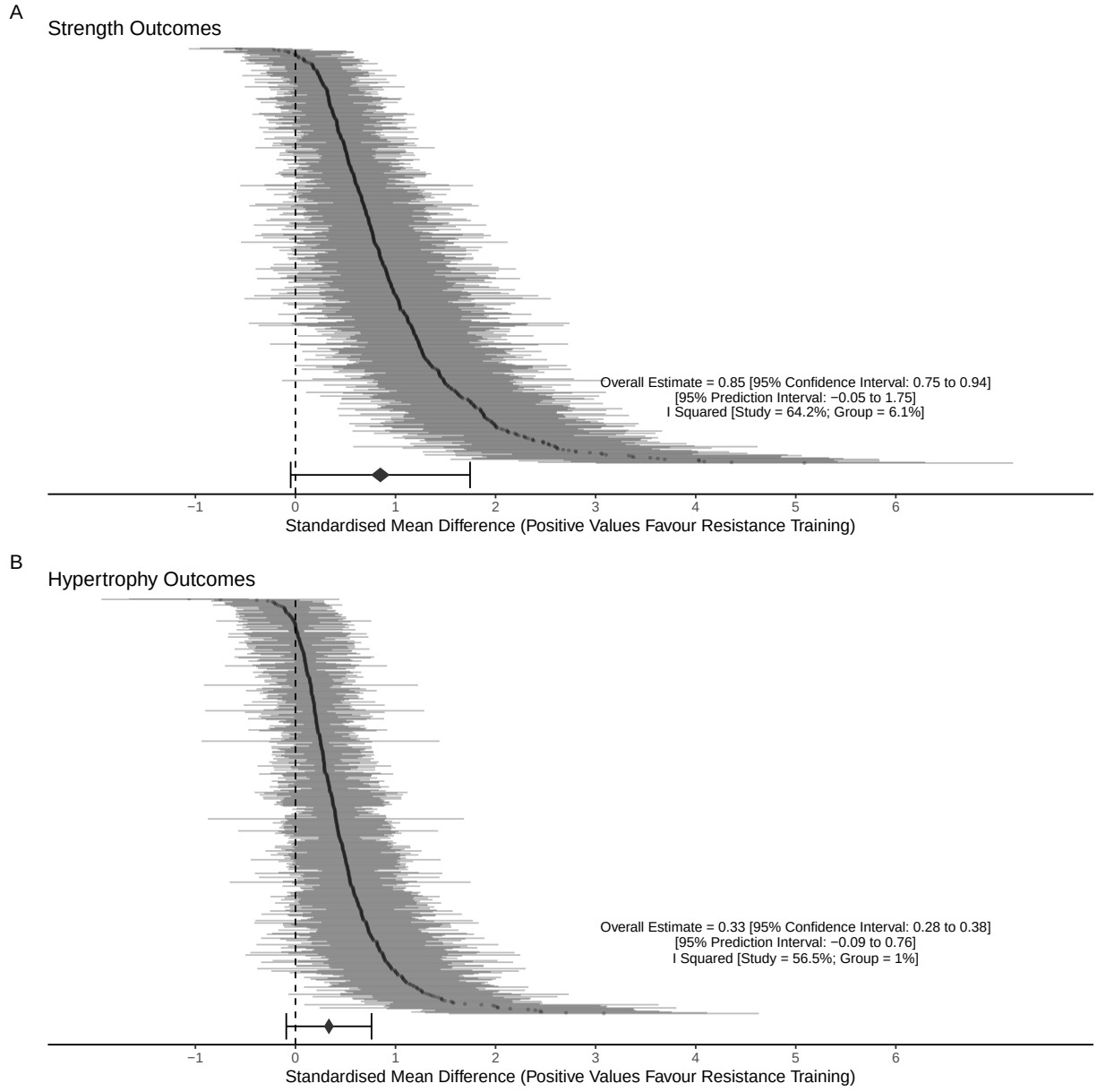

Figure 2: Caterpillar plots of SMD effect sizes for strength (A) and hypertrophy (B) outcomes.

The pattern of results from our models examining SMDs (figure 2) were similar to those reported by Polito

more influence over such biases. There did not appear to be any obvious small study bias in the dataset (see https://osf.io/stqr3).

[7]Data was coded such that study and group had explicit nesting.

et al. (2021) albeit with slightly lower estimates for both outcome types; this possibly being due to our use of a multilevel mixed-effects meta-analysis model which allowed for each effect size included to be more appropriately weighted. As might be expected, in comparison to non-training controls the RT interventions produced increases in strength (SMD = 0.85 [95%CI: 0.75 to 0.94]; I^2_{study} = 64.24, I^2_{group} = 6.08) and hypertrophy outcomes (SMD = 0.33 [95%CI: 0.28 to 0.38]; I^2_{study} = 56.48, I^2_{group} = 0.97). Confidence intervals were precise for both outcomes, though prediction intervals for SMD estimates (see figure 2) were fairly wide and relative heterogeneity was fairly high coming mostly from between study variance.

In addition to the SMD results, both the SD_{ir} and lnVR were also positive for both strength (SD_{ir} = 0.92 [95%CI: 0.36 to 1.47]; I^2_{study} = 54.02, I^2_{group} = 0; lnVR = 0.82 [95%CI: 0.64 to 1]; I^2_{study} = 87.75, I^2_{group} = 2.14) and hypertrophy outcomes (SD_{ir} = 0.33 [95%CI: 0.25 to 0.4]; I^2_{study} = 72.23, I^2_{group} = 0; lnVR = 0.46 [95%CI: 0.36 to 0.57]; I^2_{study} = 58.74, I^2_{group} = 15.14) 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. This 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* apart from the random error). However, as noted both the SD_{ir} and lnVR assume constant variance over values of the mean. As we have seen from the SMD analysis 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 apparent increase in variation. As such, we cannot rely solely on absolute comparisons of variance such as the SD_{ir} and lnVR to determine whether interindividual response variation is actually present. The lnCVR 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 and introduce a model that can also be used to overcome some possible limitations with the lnCVR.

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. We can explore the relationships between variance in a number of outcomes and the impact of certain predictors on this. For example, as noted one possible predictor of variance is the mean itself. As such, we can model variance as the response itself. The standard deviation is however bounded at zero and so in many cases it may not conform to assumptions of normality. 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) using a multilevel mixed-effects model.

As can be seen in figure 3(A) & (B), there is considerable heteroskedasticity in the relationship between the raw mean ($\overline{x}$) and standard deviation (s). 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).

$$s^2 = a\overline{x}^b \tag{15}$$

where a and b are some constants. When this relationship (equation (15)) 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:

$$2\ln s = \ln a + b\ln\overline{x} \tag{16}$$

Figure 3(C) & (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$\overline{x}$ come from outcomes over multiple groups and studies we can also estimate this relationship using the following model:

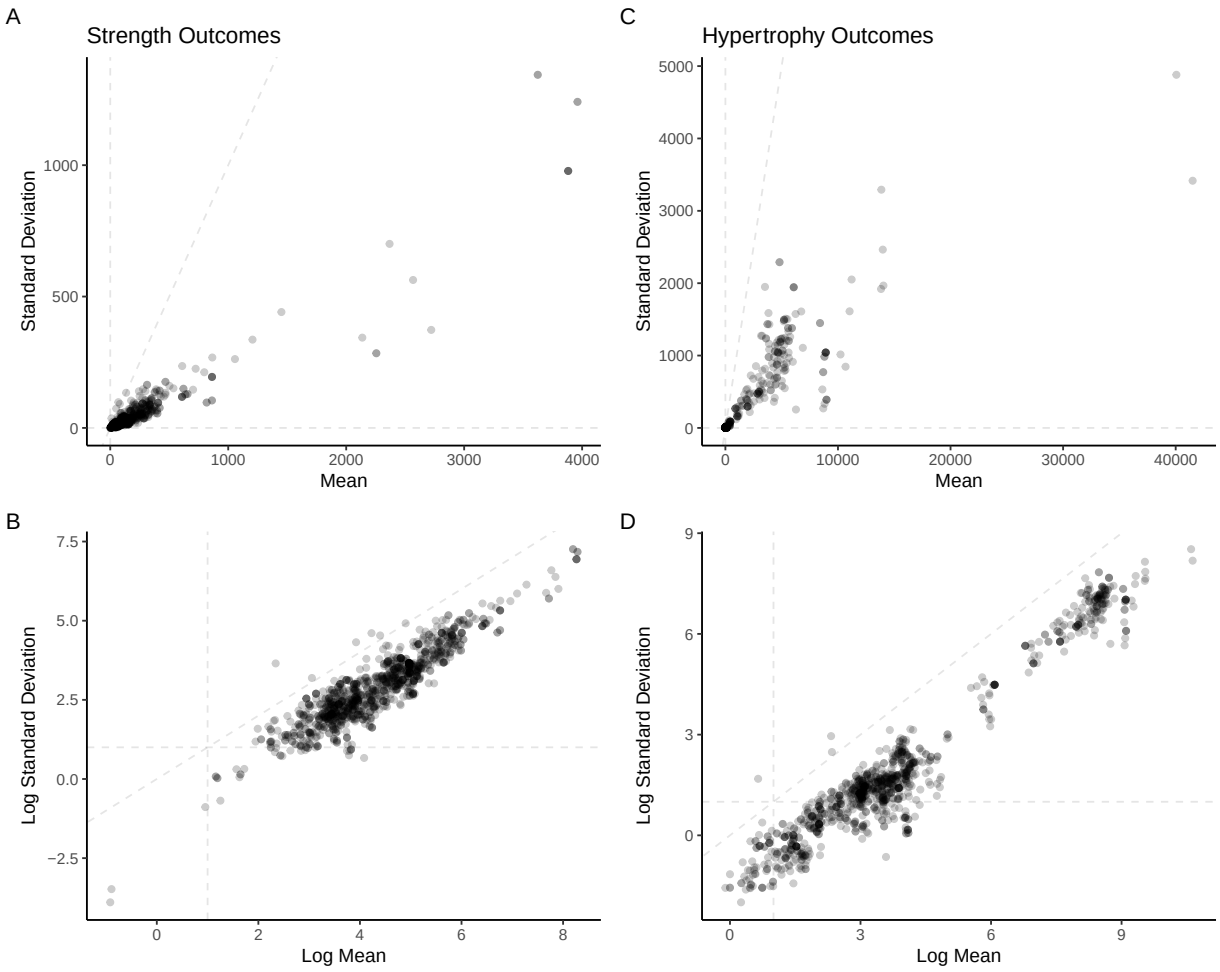

Figure 3: Scatter plots of raw mean and standard deviation of pre-intervention scores for (A) strength outcomes and (B) hypertrophy outcomes, and of the log mean and log standard deviation of pre-intervention scores for (C) strength outcomes and (D) hypertrophy outcomes.

$$\ln\hat{\sigma}_{ijk} = (\beta_0 + \tau_{(1)i} + \tau_{(2)j}) + \beta_1 \ln\overline{x}_{ijk} + \epsilon_{ijk} + m_{ijk} \tag{17}$$

where $\ln\hat{\sigma}_{ij}$ is the kth effect size, as in equation (9), from the jth group ($j = 1, 2, \cdots, N_j$; where N_j is the number of groups) in the ith study ($i = 1, 2, \cdots, N_i$; where N_i is the number of studies), $\ln\overline{x}_{ijk}$ is the mean estimate for each effect size, β_0 is the intercept, β_1 is the slope or regression coefficient for $\ln\overline{x}$, τ_i is the deviation from β_0 for the ith study and τ_j is the deviation for the jth group, and ϵ_{ijk} is the residual for each effect size which is normally distributed with σ^2_ϵ, and m_{ijk} is the sampling error for each effect size normally distributed with $\sigma^2_{\ln\hat{\sigma}_{ijk}}$. Additional predictor terms could be added to this model; for example, we could model a categorical variable for the outcome type and include $(\beta_2 + \varphi_i + \varphi_j)Outcome$ 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 slope or regression coefficient for *Outcome* (most intuitively thought of as the difference between the two outcome groups), and φ_i is the deviation (random slope) from β_2 for the ith study and φ_j is the deviation for the jth group. In this case τ_i and φ_i, and τ_j and φ_j are assumed to have multivariate normal distributions with the following variance-covariance structure:

$$\begin{pmatrix} \tau \\ \varphi \end{pmatrix} = (\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_\tau & \rho\sigma_\tau\sigma_\varphi \\ \rho\sigma_\tau\sigma_\varphi & \sigma^2_\varphi \end{pmatrix}) \tag{18}$$

Figure 4 shows this model fit visually. 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 systematically greater degree of variance in strength measures compared to hypertrophy for a given mean score.

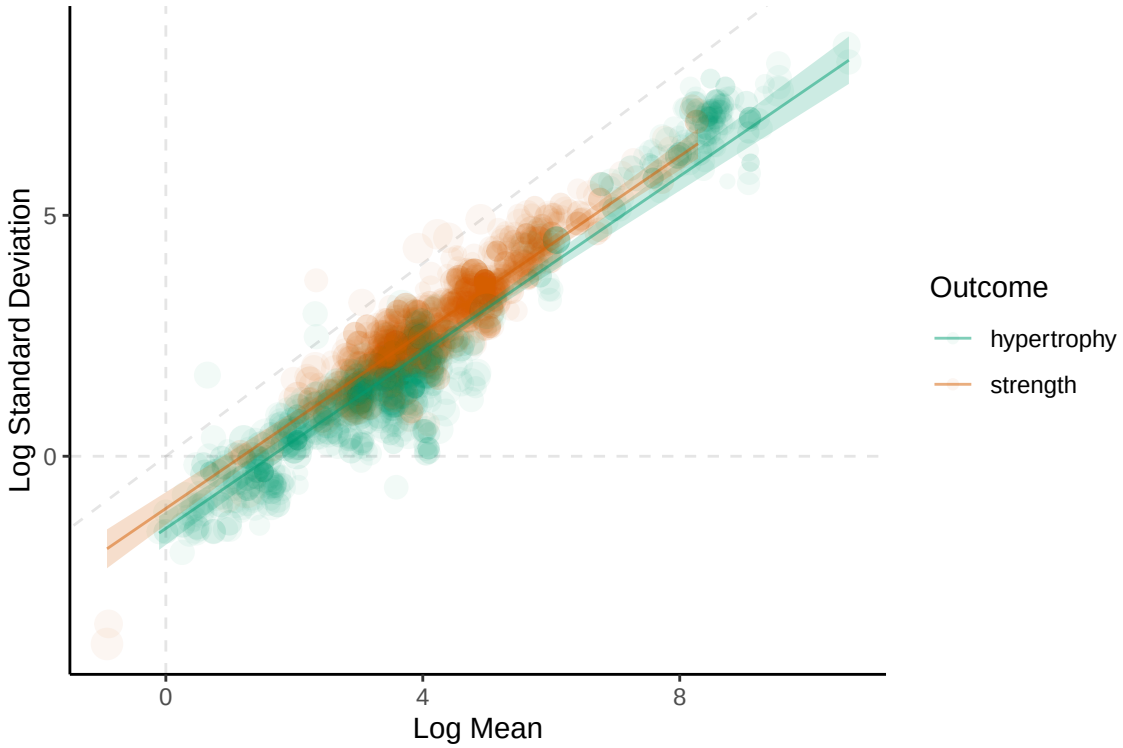

Figure 4: Meta-analytic scatter plot of the log mean and log standard deviation of pre-intervention scores.

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 lnVR) to determine whether interindividual response variation is actually present. So we should also explore this for the change-scores in the RT and control groups and determing the appropriate effects to explore.

3.3 Reanalysis of interindividual response variation using lnCVR and the random slope mixed effects model

As can be seen in figure 5 there is a mean-variance relationship in the change score data 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. As such, we cannot rely solely on absolute comparisons of variance such as the SD_{ir} and lnVR to determine whether interindividual response variation is actually present.

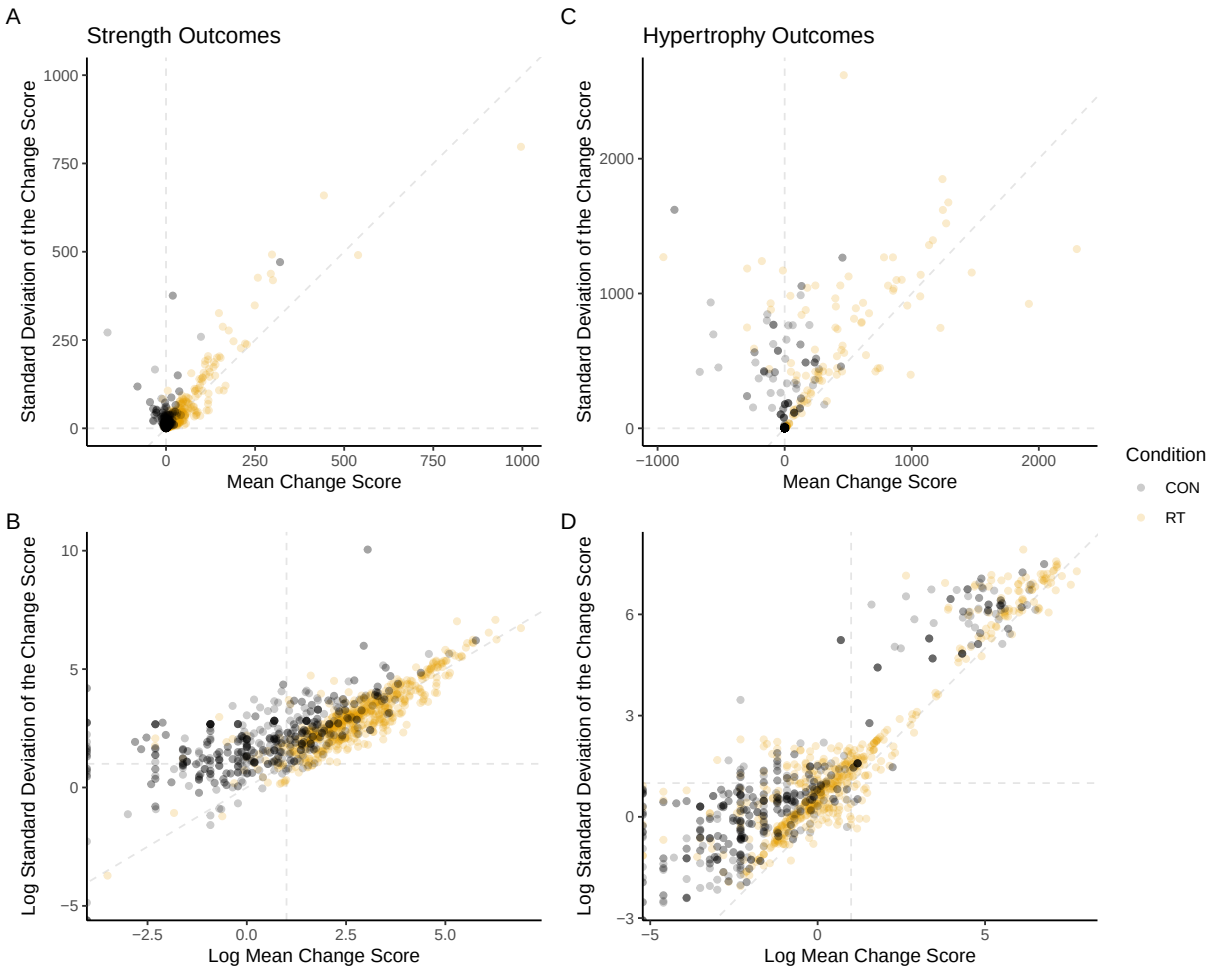

Figure 5: Scatter plots of raw mean and standard deviation of change scores for (A) strength outcomes and (B) hypertrophy outcomes, and of the log mean and log standard deviation of change scores for (C) strength outcomes and (D) hypertrophy outcomes.

The lnCVR 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 and group as before using the lnCVR as the effect size statistic leads to different conclusions compared to absolute variance comparisons. The introduction of an RT intervention actually *reduces* the relative variation seen in change scores for

strength ($\ln CVR$ = -0.62 [95%CI: -0.77 to -0.48]; I^2_{study} = 23.28, I^2_{group} = 0) and hypertrophy ($\ln CVR$ = -0.47 [95%CI: -0.62 to -0.31]; I^2_{study} = 10.33, I^2_{group} = 0) and further there is lower relative heterogeneity between studies in the effect estimates.

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 $\ln VR$). But 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 5) an alternative approach that is equivalent may be more appropriate.

The multilevel mixed-effects meta-analysis model using $\ln CVR$ as used above can be written as follows:

$$\ln CVR_{ijk} = \mu + \tau_i + \tau_j + m_{ij} \tag{19}$$

where $\ln CVR_{ij}$ is the kth effect size, as in equation (13), for the jth group ($j = 1, 2, \cdots, N_j$; where N_j is the number of groups) in the ith study ($i = 1, 2, \cdots, N_i$; where N_i is the number of studies), μ is the intercept or overall mean, τ_i is the deviation from μ for the ith study and τ_j is the deviation for the jth group, which are assumed to be normall distributed around zero with variance of σ^2_τ, and m_{ijk} is the sampling error for each effect size normally distributed with $\sigma^2_{\ln CVR_{ijk}}$. Instead, we can use a version of the random slope model described above (equation (17) and following paragraph) to compare the variability between intervention and control groups using $\ln\hat{\sigma}$ and $\ln\overline{x}$. In this case, the categorical variable for the outcome type is instead swapped for the group and the new model term included becomes $(\beta_2 + \varphi_i + \varphi_j)Group$ 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 $Group$, and φ_i is the deviation (random slope) from β_2 for the ith study and φ_j is the deviation for the jth group.

Given the heteroskedasticity in the change scores means and standard deviations (see figure 5) we fit this model to the dataset. The results were largely similar, albeit slightly attenuated, to those found using the $\ln CVR$ model for strength ($\beta_{\ln\hat{\sigma}[Group \text{ for RT}]}$ = -0.34 [95%CI: -0.57 to -0.12]) and hypertrophy ($\beta_{\ln\hat{\sigma}[Group \text{ for RT}]}$ = -0.36 [95%CI: -0.67 to -0.06]). See figure 6.

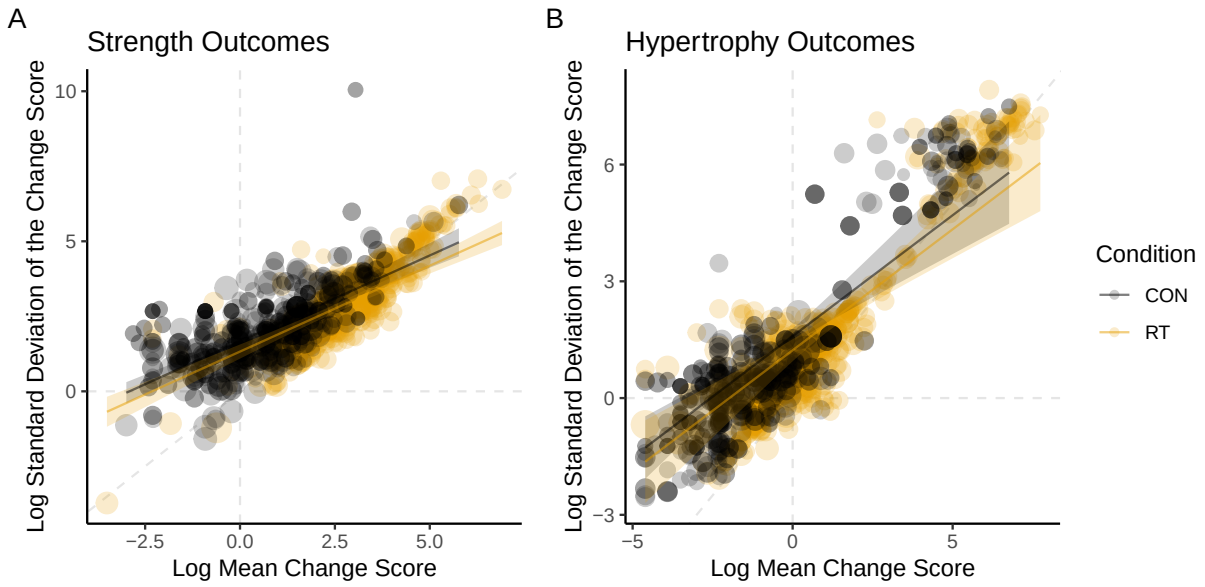

Figure 6: Meta-analytic scatter plot of the log mean and log standard deviation of change scores.

Hopefully it is clear from the regression model here where we have included both fixed and random predictors as both categorical (i.e., *Outcome*, or *Group*) and continuous (i.e., $\ln\overline{x}$) that there is considerable flexibility in the inclusion of predictors (i.e., meta-regression) when exploring variance through a meta-analytic framework. Models can be fit to explore not only how study and group level characteristics moderate effect size estimates when considering not only SMDs, but also when considering the variance based effect size statistics and models employed in this article[8].

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 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 samples sizes are typically small, meta-analysis becomes even more valuable as such small sample in primary studies have even lower statistical power to detect differences in variation as compared to means (Yang et al. 2022). The examples we presented used data from RT studies included in a recent meta-analysis published in the *Journal of Sport Sciences* (Polito, Papst, and Farinatti 2021), and so hopefully are more relatable for researchers in sport and exercise sciences.

Given the considerable interest in *precision* or *personalised* exercise prescription in particular, it is of particular interest to note the different conclusions drawn dependent upon the approach taken to determining from non-training control and RT intervention data whether or not there is *detectable* inter-individual response variation present. Using absolute comparisons of variance such as SD_{ir} and $\ln VR$ gave the impression that the introduction of the RT intervention likely increased variance above random error, suggesting the presence of inter-individual response variation. In the case of RT interventions there is evidently an average intervention effect which is positive and as such this result might merely imply that while all likely benefit, some benefit much moreso than others. Exploring variation however even in the absence of average intervention effects might still be important as those with 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 RT interventions to maximise strength and hypertrophy outcomes from our example.

However, much like that seen in cross-sectional pre-intervention scores here and most physical and biological systems, change scores demonstrated a mean-variance relationship in addition to heteroskedasticity. The likely more appropriate analyses in this case using the $\ln CVR$ in addition to the random slope model of $\ln\hat{\sigma}$ revealed conclusions in the opposite direction of the absolute variance comparisons; essentially, that the introduction of the RT intervention may have slightly *decreased* change score variance implying that there is likely little to no inter-individual response variation to explain. Interventions, such as RT interventions explored here, which induce both meaningful SMDs 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. For example, muscle strengthening interventions such as RT are recommended for *everyone* in current physical activity guidelines and in such applications there is likely value in a simple approach to recommendations (Steele et al. 2017, 2022).

The reason for this apparent reduction in variation after introduction of an RT intervention is not necessarily discernible from this analysis. Perhaps the introduction of an RT intervention has indirect effects that reduce other sources of random variance (e.g., diet, other physical activity etc.; Halliday et al. (2017)), or a ceiling effect on change (i.e., plateau in response; Steele et al. (2022)) has a constraining effect (Cortés Martínez

[8]See supplementary materials (https://osf.io/e6vpr) for model estimates for both SMD and $\ln CVR$ (used as results for $\ln CVR$ and the random slope model were similar) across a range of categorical and continous predictors for both strength and hypertrophy outcomes. There were no obvious moderators of $\ln CVR$ in particular.

2021). This potentially represents another interesting area of future study regarding variation however; 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.

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. Where considerable differences in variation are present, this suggests a meaningful line of research should be aimed at identifying subgroup- or participant-by-intervention interactions. Where variance differences are limited this suggests that translational work towards generalisable implementation might be most meaningful. Future meta-analyses in the field of sport and exercise science should consider the value of concomitantly exploring means and variances utlising the established approaches (Nakagawa et al. 2015; Hopkins 2015; Atkinson and Batterham 2015; Atkinson, Williamson, and Batterham 2019; Usui et al. 2021; Mills et al. 2021) presented here and echoing the efforts of other recent work (Kelley, Kelley, and Callahan 2022; Kelley, Kelley, and Pate 2020; Esteves et al. 2021; Bonafiglia et al. 2022; Steele et al. 2021; Fisher et al. 2022).

6 References

Atkinson, Greg, and Alan M. Batterham. 2015. "True and False Interindividual Differences in the Physiological Response to an Intervention." *Exp Physiol* 100 (6): 577–88. https://doi.org/10.1113/EP085070.

Atkinson, Greg, Philip Williamson, and Alan M. Batterham. 2019. "Issues in the Determination of 'Responders' and 'Non-Responders' in Physiological Research." *Exp Physiol* 104 (8): 1215–25. https://doi.org/10.1113/EP087712.

Bernárdez-Vázquez, Roberto, Javier Raya-González, Daniel Castillo, and Marco Beato. 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, Jacob T., Paul A. Swinton, Robert Ross, Neil M. Johannsen, Corby K. Martin, Timothy S. Church, Cris A. Slentz, et al. 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 Med*, July. https://doi.org/10.1007/s40279-022-01725-9.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to Meta-Analysis*. John Wiley & Sons.

Caldwell, Aaron, and Andrew D. Vigotsky. 2020. "A Case Against Default Effect Sizes in Sport and Exercise Science." *PeerJ* 8: e10314. https://doi.org/10.7717/peerj.10314.

Carpinelli, Ralph N. 2017. "INTERINDIVIDUAL HETEROGENEITY OF ADAPTATIONS TO RESISTANCE TRAINING." *Med Sport Pract* 18 (4): 79–94.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge. https://doi.org/10.4324/9780203771587.

Cortés Martínez, Jordi. 2021. "Constant Effect in Randomized Clinical Trials with Quantitative Outcome : A Methodological Review." *TDX (Tesis Doctorals En Xarxa)*, May. https://upcommons.upc.edu/handle/2117/349575.

Curtis, Peter S., and Xianzhong Wang. 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, Sanjoy K., Daniel R. Brown, Lewis A. Gough, Christopher P. Mclellan, Paul A. Swinton, S. Andy Sparks, and Lars R. Mcnaughton. 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): 243–56. https://doi.org/10.1080/17461391.2017.1410233.

Esteves, Gabriel Perri, Paul Swinton, Craig Sale, Ruth M. James, Guilherme Giannini Artioli, Hamilton Roschel, Bruno Gualano, Bryan Saunders, and Eimear Dolan. 2021. "Individual Participant Data Meta-

Analysis Provides No Evidence of Intervention Response Variation in Individuals Supplementing With Beta-Alanine." *Int J Sport Nutr Exerc Metab* 31 (4): 305–13. https://doi.org/10.1123/ijsnem.2021-0038.

Exner, Ryan J., Mana H. Patel, Dominic V. Whitener, Samuel L. Buckner, Matthew B. Jessee, and Scott J. Dankel. 2022. "Does Performing Resistance Exercise to Failure Homogenize the Training Stimulus by Accounting for Differences in Local Muscular Endurance?" *Eur J Sport Sci*, February, 1–10. https://doi.org/10.1080/17461391.2021.2023657.

Fisher, James, James Steele, Milo Wolf, Patroklos Androulakis Korakakis, Dave Smith, and Jürgen Giessing. 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, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research." *Educational Researcher* 5 (10): 3–8. https://doi.org/10.2307/1174772.

Hagger, Martin. 2022. "Meta-Analysis." *International Review of Sport and Exercise Psychology* 15 (1): 120–51. https://doi.org/10.1080/1750984X.2021.1966824.

Halliday, Tanya M., Jyoti Savla, Elaina L. Marinik, Valisa E. Hedrick, Richard A. Winett, and Brenda M. Davy. 2017. "Resistance Training Is Associated with Spontaneous Changes in Aerobic Physical Activity but Not Overall Diet Quality in Adults with Prediabetes." *Physiology & Behavior* 177 (August): 49–56. https://doi.org/10.1016/j.physbeh.2017.04.013.

Halperin, Israel, Tomer Malleron, Itai Har-Nir, Patroklos Androulakis-Korakakis, Milo Wolf, James Fisher, and James Steele. 2022. "Accuracy in Predicting Repetitions to Task Failure in Resistance Exercise: A Scoping Review and Exploratory Meta-Analysis." *Sports Med* 52 (2): 377–90. https://doi.org/10.1007/s40279-021-01559-x.

Hecksteden, Anne, Jochen Kraushaar, Friederike Scharhag-Rosenberger, Daniel Theisen, Stephen Senn, and Tim Meyer. 2015. "Individual Response to Exercise Training - a Statistical Perspective." *J Appl Physiol (1985)* 118 (12): 1450–59. https://doi.org/10.1152/japplphysiol.00714.2014.

Hedges, L. V., and A. Nowell. 1995. "Sex Differences in Mental Test Scores, Variability, and Numbers of High-Scoring Individuals." *Science* 269 (5220): 41–45. https://doi.org/10.1126/science.7604277.

Hedges, Larry V., Jessica Gurevitch, and Peter S. Curtis. 1999. "The Meta-Analysis of Response Ratios in Experimental Ecology." *Ecology* 80 (4): 1150–56. https://doi.org/10.1890/0012-9658(1999)080%5B1150:TMAORR%5D2.0.CO;2.

Hedges, Larry V., and Ingram Olkin. 2014. *Statistical Methods for Meta-Analysis*. Academic Press.

Heidel, Kyle A., Zachary J. Novak, and Scott J. Dankel. 2022. "Machines and Free Weight Exercises: A Systematic Review and Meta-Analysis Comparing Changes in Muscle Size, Strength, and Power." *J Sports Med Phys Fitness* 62 (8): 1061–70. https://doi.org/10.23736/S0022-4707.21.12929-9.

Hopkins, Will G. 2015. "Individual Responses Made Easy." *J Appl Physiol (1985)* 118 (12): 1444–46. https://doi.org/10.1152/japplphysiol.00098.2015.

Hrubeniuk, Travis J., Jacob T. Bonafiglia, Danielle R. Bouchard, Brendon J. Gurd, and Martin Sénéchal. 2022. "Directions for Exercise Treatment Response Heterogeneity and Individual Response Research." *Int J Sports Med* 43 (1): 11–22. https://doi.org/10.1055/a-1548-7026.

Kelley, George A. 2022. "Precision Exercise Medicine in Rheumatology: Don't Put the Cart Before the Horse." *Clin Rheumatol* 41 (8): 2277–79. https://doi.org/10.1007/s10067-022-06260-6.

Kelley, George A., Kristi S. Kelley, and Leigh F. Callahan. 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." *Arch Phys Med Rehabil*, January, S0003-9993(22)00007-7. https://doi.org/10.1016/j.apmr.2021.12.019.

Kelley, George A., Kristi S. Kelley, and Russell R. Pate. 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." *Child Obes* 16 (5): 301–6. https://doi.org/10.1089/chi.2020.0056.

Lajeunesse, Marc J. 2015. "Bias and Correction for the Log Response Ratio in Ecological Meta-Analysis." *Ecology* 96 (8): 2056–63. https://doi.org/10.1890/14-2402.1.

Mills, Harriet L., Julian P. T. Higgins, Richard W. Morris, David Kessler, Jon Heron, Nicola Wiles, George Davey Smith, and Kate Tilling. 2021. "Detecting Heterogeneity of Intervention Effects Using Analysis and Meta-Analysis of Differences in Variance Between Trial Arms." *Epidemiology* 32 (6): 846–54.

https://doi.org/10.1097/EDE.0000000000001401.

Nakagawa, Shinichi, and Innes C. Cuthill. 2007. "Effect Size, Confidence Interval and Statistical Significance: A Practical Guide for Biologists." *Biol Rev Camb Philos Soc* 82 (4): 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x.

Nakagawa, Shinichi, Robert Poulin, Kerrie Mengersen, Klaus Reinhold, Leif Engqvist, Malgorzata Lagisz, and Alistair M. Senior. 2015. "Meta-Analysis of Variation: Ecological and Evolutionary Applications and Beyond." *Methods in Ecology and Evolution* 6 (2): 143–52. https://doi.org/10.1111/2041-210X.12309.

Nakagawa, Shinichi, and Holger Schielzeth. 2012. "The Mean Strikes Back: Mean-Variance Relationships and Heteroscedasticity." *Trends Ecol Evol* 27 (9): 474-475; author reply 475-476. https://doi.org/10.1016/j.tree.2012.04.003.

Pickering, Craig, and John Kiely. 2019. "Do Non-Responders to Exercise Exist-and If So, What Should We Do About Them?" *Sports Med* 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): 130–35. https://doi.org/10.2307/2333051.

Polito, Marcos D., Rafael R. Papst, and Paulo Farinatti. 2021. "Moderators of Strength Gains and Hypertrophy in Resistance Training: A Systematic Review and Meta-Analysis." *Journal of Sports Sciences* 39 (19): 2189–98. https://doi.org/10.1080/02640414.2021.1924978.

Ross, Robert, Bret H. Goodpaster, Lauren G. Koch, Mark A. Sarzynski, Wendy M. Kohrt, Neil M. Johannsen, James S. Skinner, et al. 2019. "Precision Exercise Medicine: Understanding Exercise Response Variability." *Br J Sports Med* 53 (18): 1141–53. https://doi.org/10.1136/bjsports-2018-100328.

Steele, James, James P. Fisher, Jurgen Giessing, Patroklos Androulakis-Korakakis, Milo Wolf, Bram Kroeske, and Rob Reuters. 2022. "Long-Term Time-Course of Strength Adaptation to Minimal Dose Resistance Training Through Retrospective Longitudinal Growth Modeling." *Research Quarterly for Exercise and Sport* 0 (0): 1–18. https://doi.org/10.1080/02701367.2022.2070592.

Steele, James, James Fisher, Martin Skivington, Chris Dunn, Josh Arnold, Garry Tew, Alan M. Batterham, et al. 2017. "A Higher Effort-Based Paradigm in Physical Activity and Exercise for Public Health: Making the Case for a Greater Emphasis on Resistance Training." *BMC Public Health* 17 (1): 300. https://doi.org/10.1186/s12889-017-4209-8.

Steele, James, Daniel Plotkin, Derrick Van Every, Avery Rosa, Hugo Zambrano, Benjiman Mendelovits, Mariella Carrasquillo-Mercado, Jozo Grgic, and Brad J. Schoenfeld. 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)* 9 (11): 155. https://doi.org/10.3390/sports9110155.

Swinton, Paul A., Ben Stephens Hemingway, Bryan Saunders, Bruno Gualano, and Eimear Dolan. 2018. "A Statistical Framework to Interpret Individual Response to Intervention: Paving the Way for Personalized Nutrition and Exercise Prescription." *Front Nutr* 5: 41. https://doi.org/10.3389/fnut.2018.00041.

Taylor, L. R. 1961. "Aggregation, Variance and the Mean." *Nature* 189 (4766): 732–35. https://doi.org/10.1038/189732a0.

Tenan, Matthew S., Andrew D. Vigotsky, and Aaron R. Caldwell. 2020. "Comment on: 'A Method to Stop Analyzing Random Error and Start Analyzing Differential Responders to Exercise'." *Sports Med* 50 (2): 431–34. https://doi.org/10.1007/s40279-019-01249-9.

Usui, Takuji, Malcolm R. Macleod, Sarah K. McCann, Alistair M. Senior, and Shinichi Nakagawa. 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.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the Metafor Package." *Journal of Statistical Software* 36 (August): 1–48. https://doi.org/10.18637/jss.v036.i03.

Vigotsky, Andrew, Gregory Lee Nuckols, James Fisher, James Heathers, James Krieger, Brad Jon Schoenfeld, Jürgen Giessing, and James Steele. 2020. "Improbable Data Patterns in the Work of Barbalho Et Al." SportRxiv. https://doi.org/10.31236/osf.io/sg3wm.

Yang, Yefeng, Helmut Hillebrand, Malgorzata Lagisz, Ian Cleasby, and Shinichi Nakagawa. 2022. "Low Statistical Power and Overestimated Anthropogenic Impacts, Exacerbated by Publication Bias, Dominate Field Studies in Global Change Biology." *Global Change Biology* 28 (3): 969–89. https://doi.org/10.1111/gcb.15972.