Meta-Analysis of Variation in Sport and Exercise Science

Examples of Application Within Resistance Training Research

????

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 well known that extcolor{red}{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, extcolor{red}{in} a field with small samples such as sport and exercise science, exploration of variance through a meta-analytic framework is appealing. extcolor{red}{Despite} the value of embracing and exploring variation alongside mean effects in sport and exercise science extcolor{red}{it is rarely applied to research synthesis through meta-analysis}. extcolor{red}{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__studies__1958?**), the modern-day term "meta-analysis" was coined by Gene Glass in (**glass__primary__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__meta-analysis__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.

As with many other fields (**nakagawa__meta-analysis__2015?**; **usui__meta-analysis__2021?**; **mills__detecting__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(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 (**bernardez-vazquez__resistance__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__case__2020?**), the standardised mean difference (SMD), is used to compare means between groups or conditions. This statistic is usually Cohen's d (**cohen__statistical__1988?**), or its bias-corrected metric referred to as Hedges' g (**hedges__statistical__2014?**; **borenstein__introduction__2021?**; **nakagawa__effect__2007?**) [2]. The SMD, and its sampling variance, s^2_{SMD} are given by:

[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__machines__2022?**)).

[2]We 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.

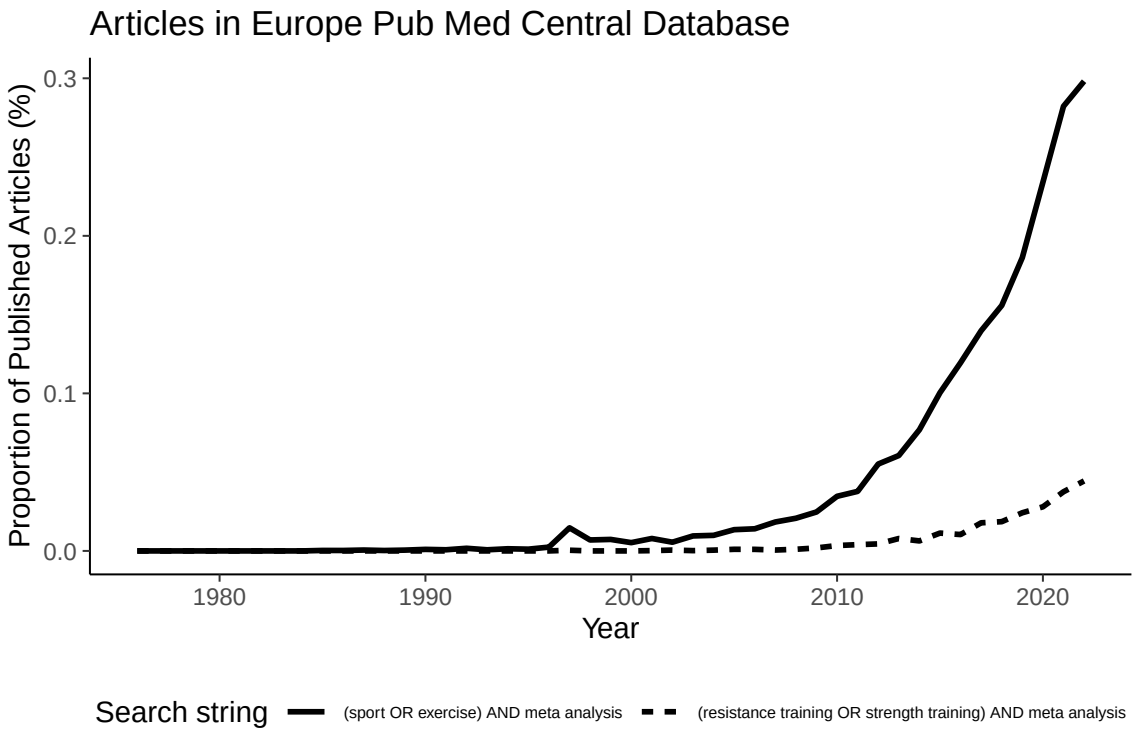

Figure 1: Trends in meta-analyses published in sport and exercise science since 1976.

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

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

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

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

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__meta-analysis__1998?**; **hedges__meta-analysis__1999?**; **lajeunesse__bias__2015?**). The lnRR, and its sampling variance, $s^2_{\ln RR}$ are given by:

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

$$s^2_{\ln RR} = \frac{s_C^2}{n_C \overline{x}_C^2} + \frac{s_C^4}{2n_C \overline{x}_C^2 \overline{x}_C^4} + \frac{s_E^2}{n_E \overline{x}_E^2} + \frac{s_E^4}{2n_E \overline{x}_E^2 \overline{x}_E^4} \tag{6}$$

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 contrast, the lnRR is uninfluenced by the standard deviations in either groups (see equation (5)), which only affects the sampling variance (see equation (6)). Despite this, the use of effects sizes like the lnRR has been limited in previous meta-analyses in sport and exercise science (**deb_quantifying_2018?**; **nuzzo_how_2022?**) and to our knowledge only one meta-analysis in RT has used this kind of effect size (**swinton_interpreting_2022?**).

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 (**hecksteden_individual_2015?**; **atkinson_true_2015?**; **atkinson_issues_2019?**; **ross_precision_2019?**; **swinton_statistical_2018?**; **hopkins_individual_2015?**; **kelley_precision_2022?**; **hrubeniuk_directions_2022?**; **pickering_non-responders_2019?**). However, despite the availability of approaches to compare variances between groups, in sport and exercise science this is rarely explored in primary research (**bonafiglia_interindividual_2022?**). Moreover, although 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_are_2022?**; **kelley_are_2020?**; **esteves_individual_2021?**; **bonafiglia_interindividual_2022?**; **steele_slow_2021?**; **fisher_role_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, typically with large samples (**hecksteden_individual_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_mean_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_does_2022?**) as well as a meta-analysis exploring between-participant heterogeneity of accuracy in predicting proximity to task failure during RT (**halperin_accuracy_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. 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_low_2022?**).

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_individual_2015?**); (**atkinson_true_2015?**); (**atkinson_issues_2019?**)), the log ratio of standard deviations (lnVR; termed the "variability ratio"; (**hedges_sex_1995?**)), and the log ratio of coefficient of variation (lnCVR; termed the "coefficient of variation ratio"; (**nakagawa_meta-analysis_2015?**); (**senior_revisiting_2020?**)). We present the

[3]Exploration of methodological approaches and their impact on heterogeneity have also been explored in preclinical research (**usui_meta-analysis_2021?**).

independent groups versions due to use of randomised controlled trials in our examples below, but note that dependent versions also exist for lnVR and lnCVR.

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 (**hopkins_individual_2015?**; **atkinson_true_2015?**; **atkinson_issues_2019?**). 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 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_individual_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 error[4] (**hecksteden_individual_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, $s^2_{SD_{ir}}$ are given by:

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

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

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 lnVR and lnCVR.

2.2 Log ratio of standard deviations (lnVR)

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 lnVR (**nakagawa_meta-analysis_2015?**; **senior_revisiting_2020?**). An unbiased estimator of the natural logarithm of a population standard deviation (lnσ), 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 lnσ, and it is assumed with sufficiently large sample size and value of σ that lnσ 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 and intervention, the lnVR, and its sampling variance, $s^2_{\ln VR}$ is given by:

[4]Though 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.

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

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

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[5] and we 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_comment_2020?**).

2.3 Log ratio of coefficient of variation ($\ln CVR$)

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 $\ln CVR$ is thus a more generally applicable effect size statistic for examining variability between groups. Considering equations (5) and (11), the $\ln CVR$ is given by:

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

where CV_E and CV_C are $s_E/\overline{x}_E$ and $s_C/\overline{x}_C$ respectively. Senior et al. (**senior_revisiting_2020?**) derived the sampling variance, $s^2_{\ln CVR}$, as:

$$\begin{aligned} s^2_{\ln CVR} = &\frac{s^2_C}{n_C \overline{x}^2_C} + \frac{s^4_C}{2n^2_C \overline{x}^4_C} + \frac{n_C}{(n_C - 1)^2} \\ &+ \frac{s^2_E}{n_E \overline{x}^2_E} + \frac{s^4_E}{2n^2_E \overline{x}^4_E} + \frac{n_E}{(n_E - 1)^2} \end{aligned} \tag{14}$$

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_moderators_2021?**). Here we have used their list of included studies and re-extracted data from 111 of these[6]. 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_conducting_2010?**). The extracted dataset, analysis scripts, models, data summaries, and supplementary materials are available on the Open Science Framework (https://osf.io/2h9ma/).

Polito et al. (**polito_moderators_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,

[5]For one clear example, see figure 1A in Vigostky et al. (**vigotsky_improbable_2020?**) who show that the mean and standard deviation for baseline strength values typically scale with one another across most studies.

[6]The 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).

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

with both overall effect estimate and moderator analyses (i.e., meta-regressions) performed. Given that Polito et al. (**polito_moderators_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 means of 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 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 variance (**yang_low_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

First we conducted a traditional SMD and lnRR based effect size[7] meta-analysis to explore the effects of RT interventions compared to controls for strength and hypertrophy (i.e., muscle mass/size) outcomes[8]. Polito et al. (**polito_moderators_2021?**) originally used a normal random-effects meta-analysis. However, the data we extracted were hierarchical in nature (multiple outcomes measures within each arm, i.e., intervention group(s) and control group, within each study) and thus a multilevel mixed-effects meta-analysis model (**van_den_noortgate_three-level_2013?**) with cluster-robust variance estimation (**hedges_robust_2010?**) was used with random intercepts for study, arm[9], and effect. We then fitted the same model for the SD_{ir} and lnVR 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 lnRR would indicate that RT interventions produced greater improvements in outcomes

[7]It 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-postest-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 lnRR effect sizes in equations (1) and (5) 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., (**gurevitch_interaction_2000?**); Morris et al., (**morris_direct_2007?**); Morris (**morris_estimating_2008?**); Lajeunesse (**lajeunesse_bias_2015?**)) and these are the effect sizes used in the meta-analyses referred to here.

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

[9]We use the term *arm* to refer to an intervention group-control group contrast to accomodate 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, lnRR, SD_{ir}, lnVR, and lnCVR) 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 + 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)

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).

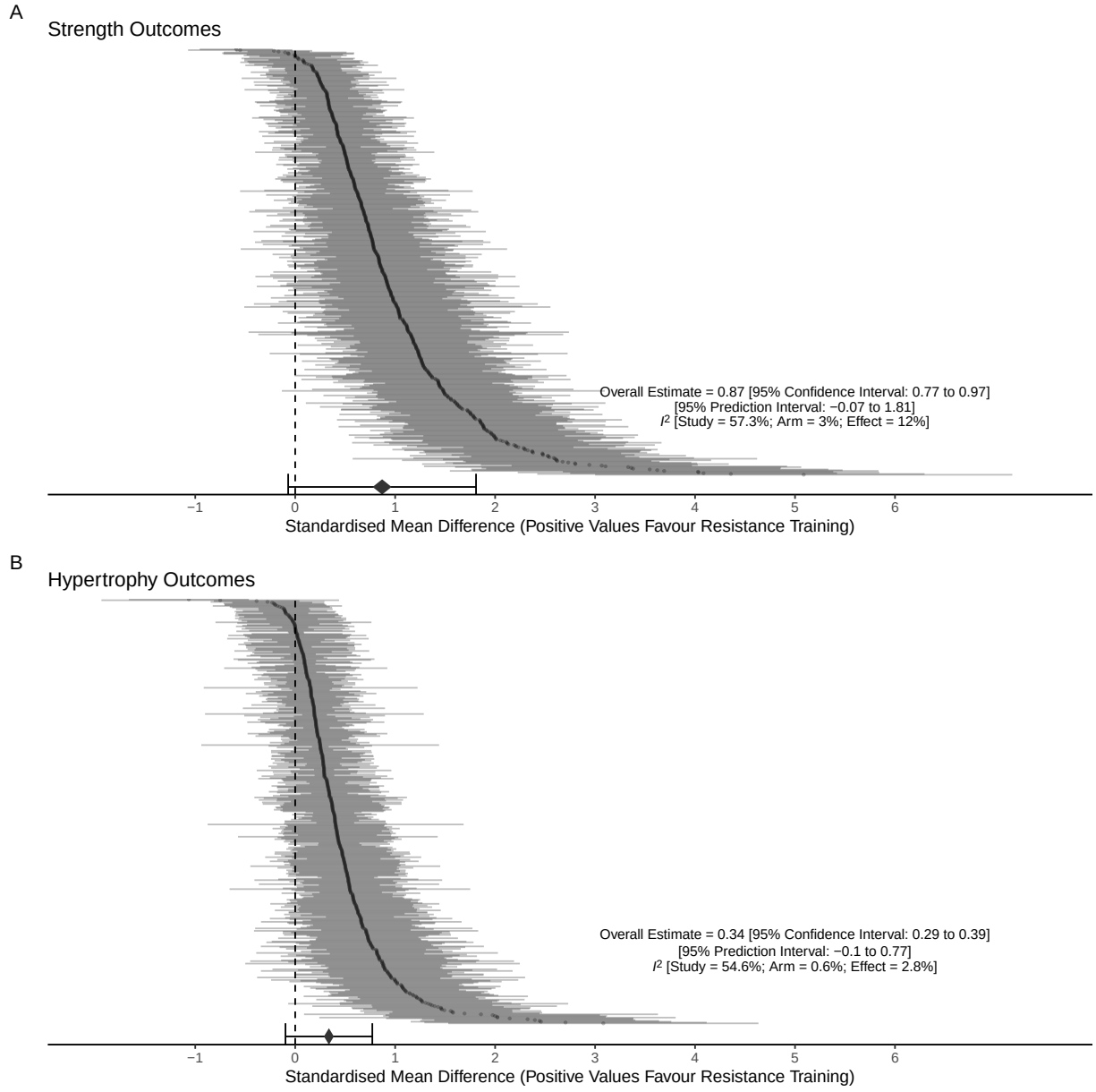

Figure 2: Caterpillar plots of SMD effect sizes for strength (A) and hypertrophy (B) outcomes.

The pattern of results from our models examining SMDs (figure 2) were similar to those reported by Polito et al. (**polito_moderators_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. 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^2_{study} = 57.32%, I^2_{arm} = 3%, I^2_{effect} = 11.95%) and hypertrophy outcomes (SMD = 0.34 [95%CI: 0.29 to 0.39]; I^2_{study} = 54.62%, I^2_{arm} = 0.62%, I^2_{effect} = 2.79%). 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 mostly

as a result of between-study variance.

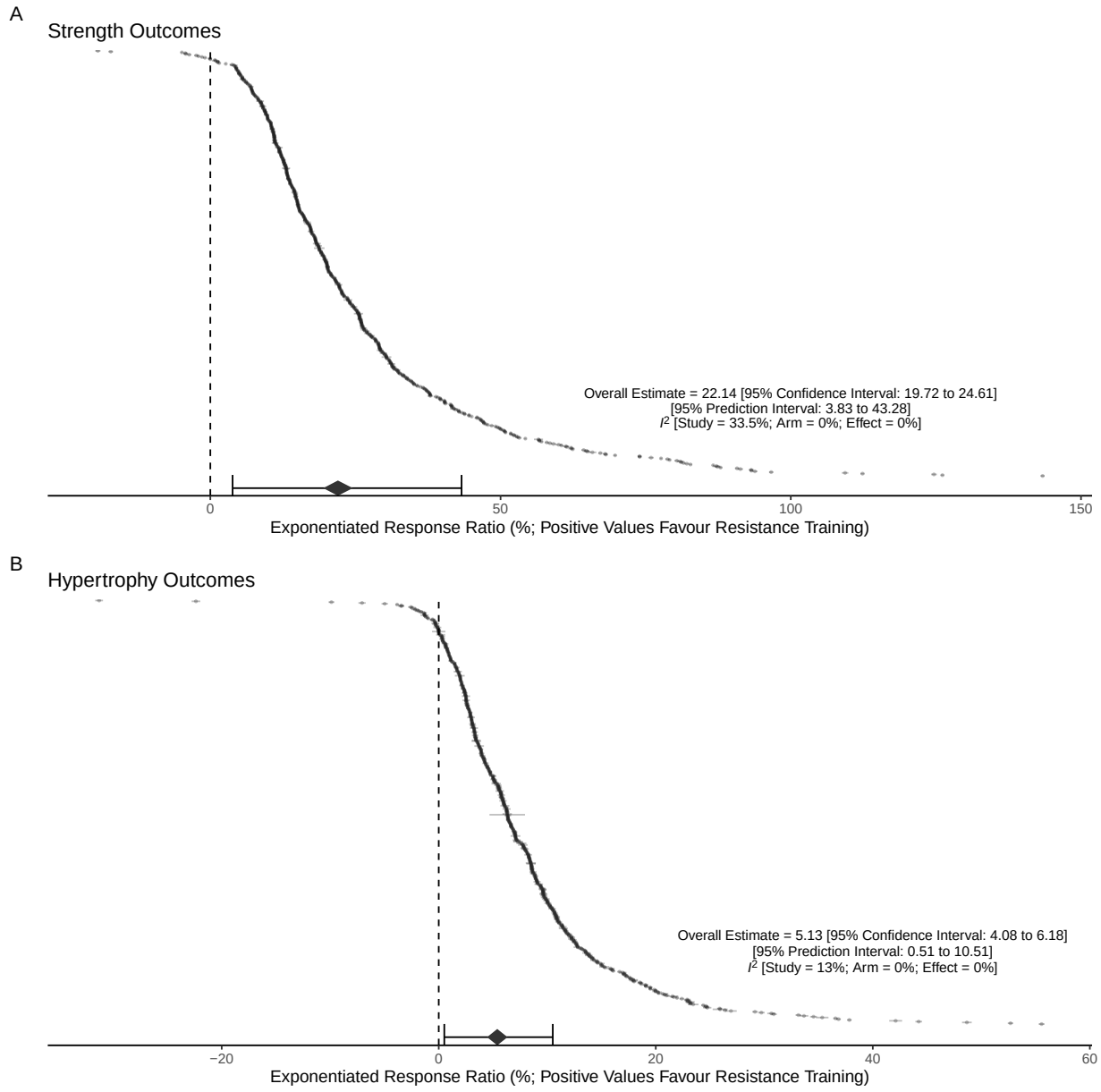

Figure 3: Caterpillar plots of exponentiated RR effect sizes for strength (A) and hypertrophy (B) outcomes.

For the lnRR 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 (expRR = 21.97 [95%CI: 19.43 to 24.57]; I^2_{study} = 33.46%, I^2_{arm} = 0%, I^2_{effect} = 0%) and hypertrophy (expRR = 5.39 [95%CI: 4.44 to 6.35]; I^2_{study} = 12.97%, I^2_{arm} = 0%, I^2_{effect} = 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 lnRR results, both the SD_{ir} (figure 4) and lnVR (figure 5) were also positive for both strength (SD_{ir} = 0.91 [95%CI: 0.36 to 1.47]; I^2_{study} = 53.85%, I^2_{arm} = 0.04%, I^2_{effect} = 0%; lnVR = 0.9 [95%CI: 0.77 to 1.02]; I^2_{study} = 56.36%, I^2_{arm} = 0.77%, I^2_{effect} = 27.36%) and hypertrophy outcomes (SD_{ir}

= 0.42 [95%CI: 0.33 to 0.5]; $I^2_{study} = 0.01\%$, $I^2_{arm} = 40.15\%$, $I^2_{effect} = 39.73\%$; $\ln VR = 0.5$ [95%CI: 0.4 to 0.6]; $I^2_{study} = 41.21\%$, $I^2_{arm} = 3.31\%$, $I^2_{effect} = 33.59\%$) 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_interindividual_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. 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 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 and introduce a model that can also be used to overcome some possible limitations with the $\ln CVR$.

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. (**polito_moderators_2021?**).

As can be seen in figure 6(A) and (C), 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_aggregation_1961?**).

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

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:

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

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 $\ln\overline{x}$ come from multiple outcomes within multiple arms within studies we can also estimate this relationship using a multilevel mixed-effects meta-regression model. For example, the following model specifies $\ln\overline{x}$ as a fixed effect with random intercepts for study and arm:

$$\ln\hat{\sigma}_{ijk} = (\beta_0 + \tau_{(1)i} + \tau_{(2)j} + \tau_{(3)k}) + \beta_1\ln\overline{x}_{ijk} + \epsilon_{ijk} + m_{ijk} \tag{17}$$

where $\ln\hat{\sigma}_{ijk}$ is the kth effect size, as in equation (9), from the jth arm ($j = 1, 2, \cdots, N_j$; where N_j is the

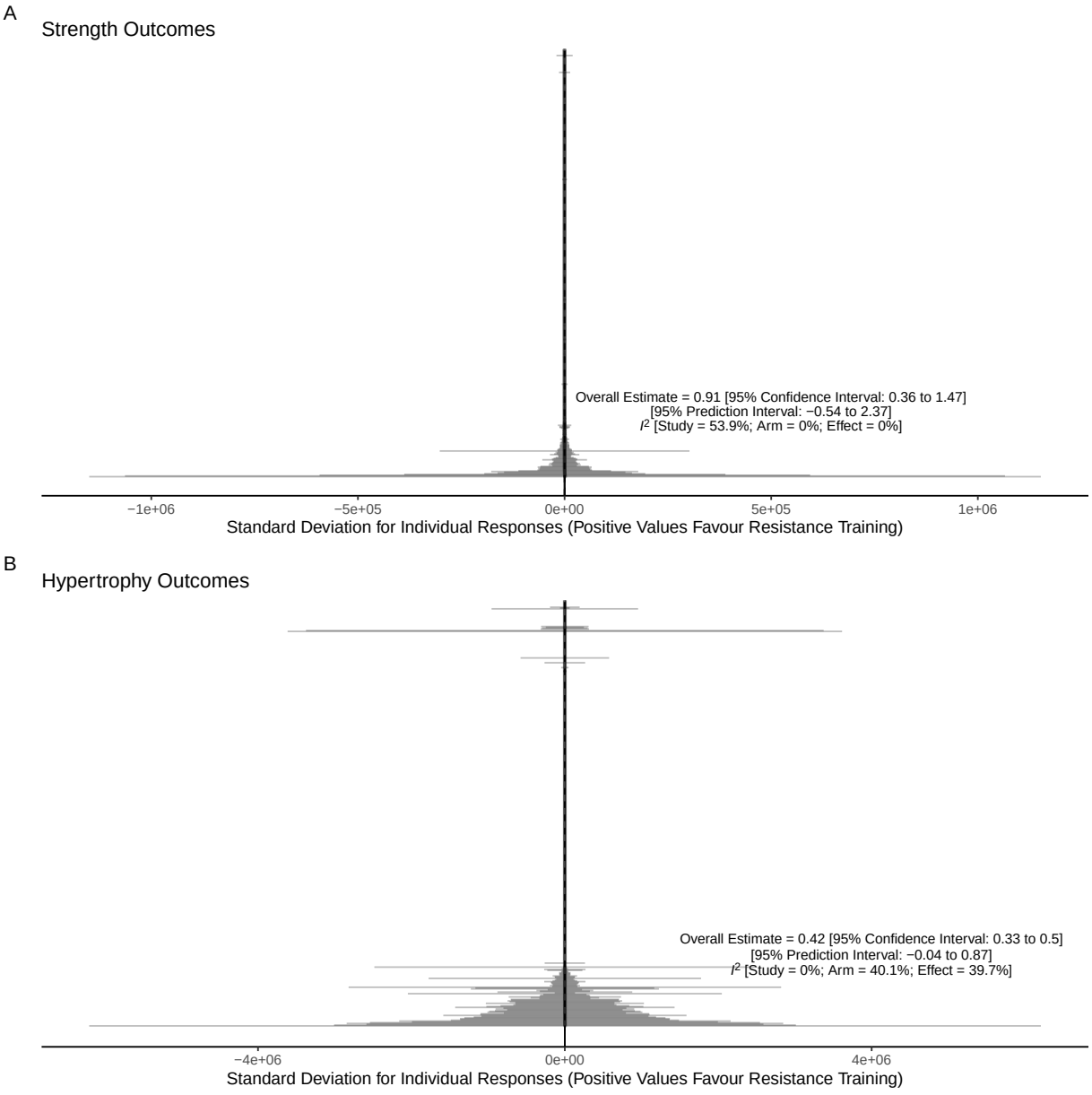

Figure 4: Caterpillar plots of SDir effect sizes for strength (A) and hypertrophy (B) outcomes.

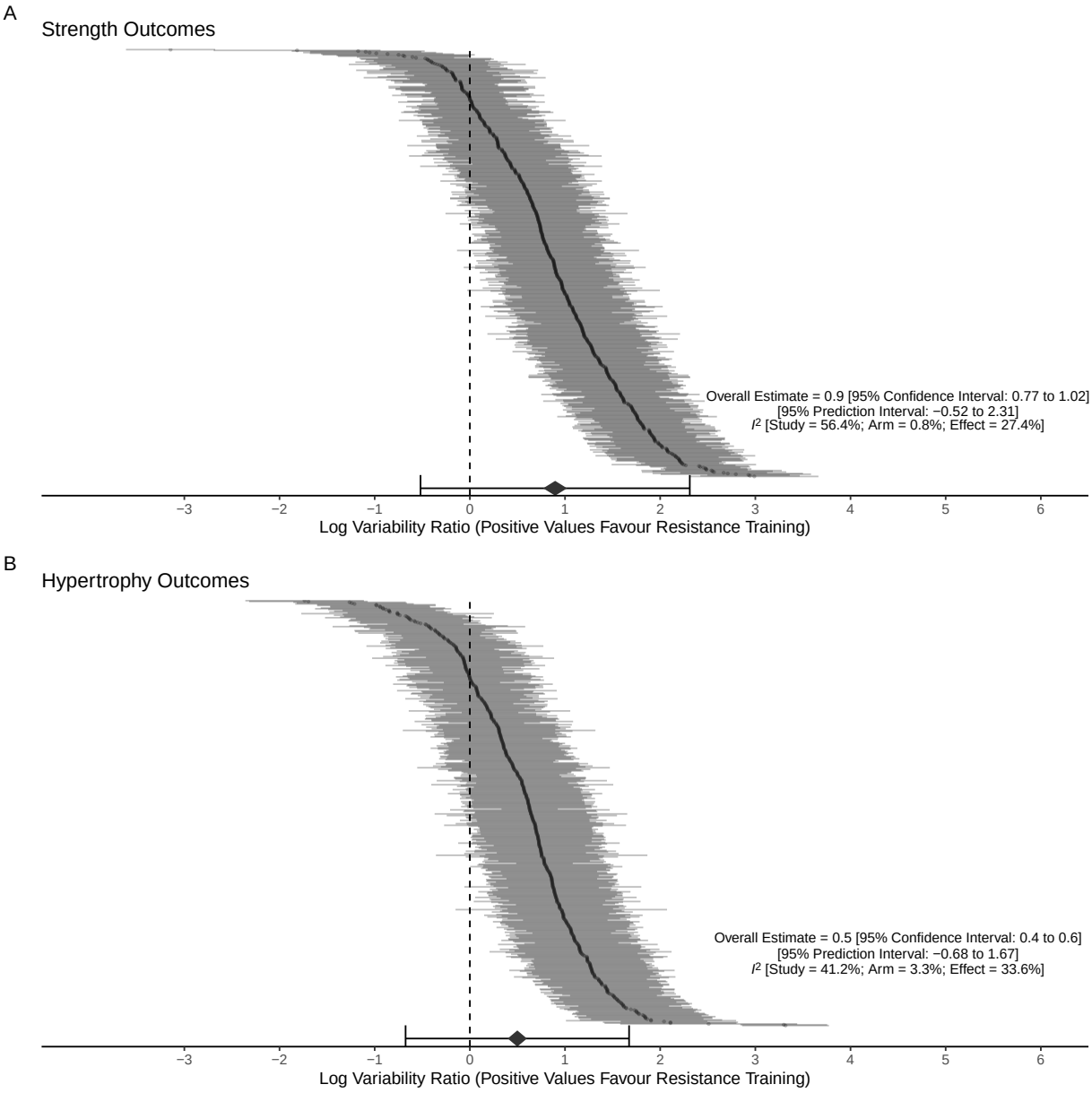

Figure 5: Caterpillar plots of $\ln VR$ effect sizes for strength (A) and hypertrophy (B) outcomes.

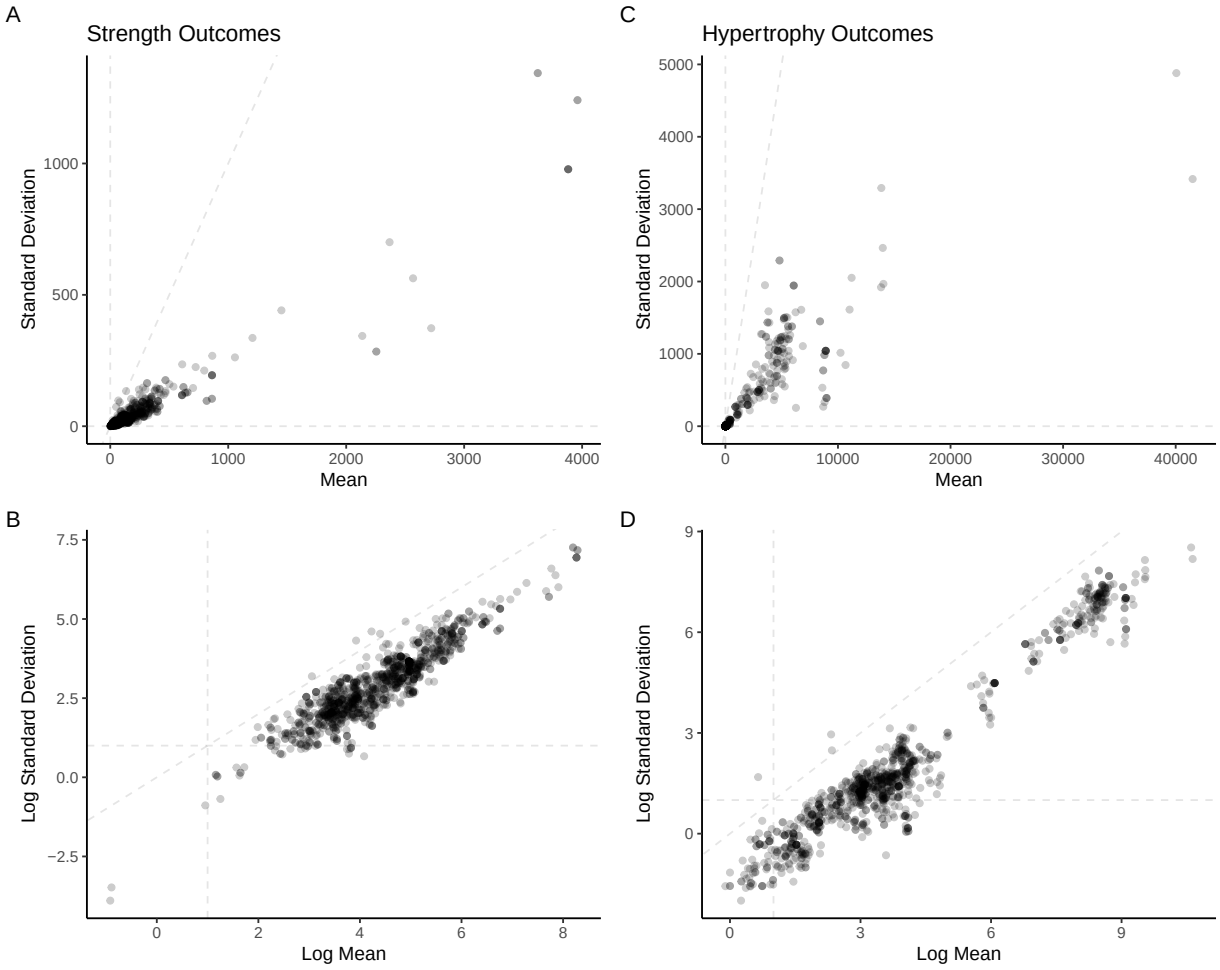

Figure 6: Scatter plots of raw mean and standard deviation of pre-intervention scores for (A) strength outcomes and (B) hypertrophy outcomes, and of the log mean and log standard deviation of pre-intervention scores for (C) strength outcomes and (D) hypertrophy outcomes.

number of arms[10]) 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, τ_j is the deviation for the jth arm, τ_k is the deviation for the kth effect, 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 $\beta_2 Outcome$ 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)[11].

Figure 7 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 but systematically greater degree of variance in strength measures compared to hypertrophy for a given mean score.

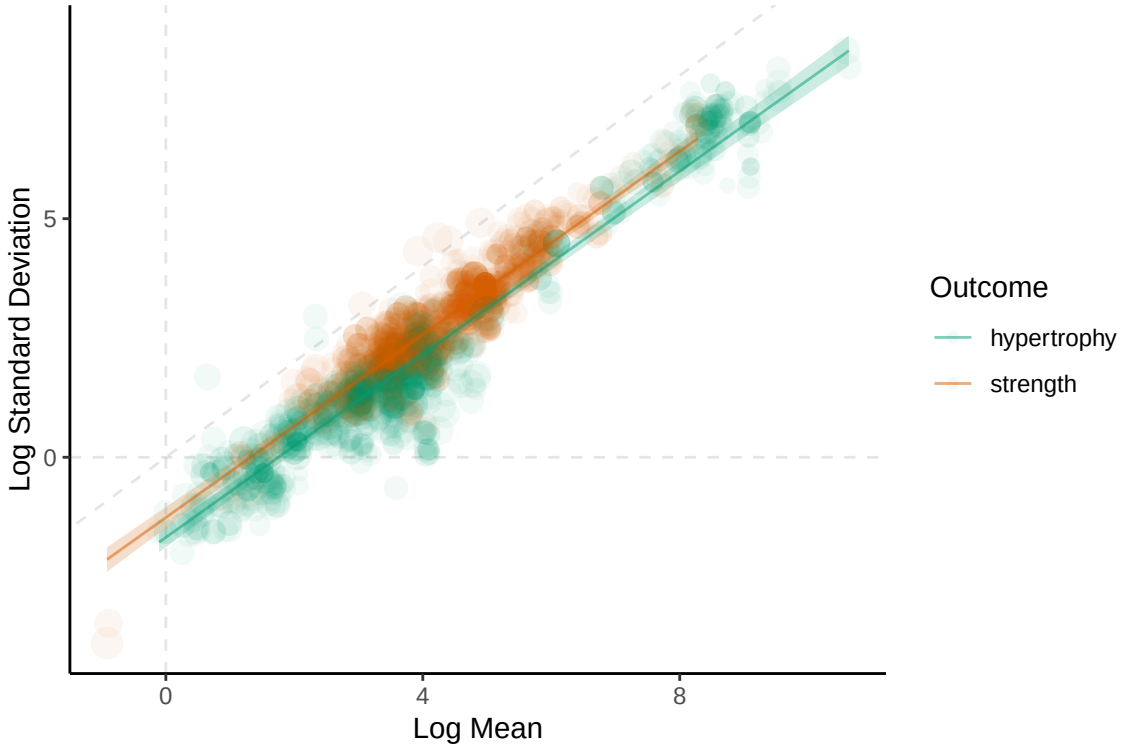

Figure 7: Meta-analytic scatter plot of the log mean and log standard deviation of pre-intervention scores.

The presence of Taylor's law type relationships should be examined in datasets prior to deciding on which

[10]In contrast to the models examining effect sizes relating to comparisons between an intervention group and control group, in the models examining $\ln\hat{\sigma}$ with $\ln\overline{x}$ as a predictor the term *arm* refers to both intervention group(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 was again coded such that study and arm had explicit nesting.

[11]We 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 to enhance generalisability of inferences (**barr_random_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 slope or regression coefficient for $Outcome$, and φ_i is the deviation (random slope) from β_2 for the ith study and φ_j is the deviation for the jth arm. Indeed, we fit additional models using $\ln\hat{\sigma}$ with $\ln\overline{x}$ and $Outcome$ as a predictor with (1) random intercepts for study and arm only, (2) the inclusion of random slopes for $\ln\overline{x}$ by study, and (3) the inclusion of random slopes for $\ln\overline{x}$ 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.

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 for the change-scores in the RT and control groups and determining the appropriate effects to explore.

3.3 Reanalysis of interindividual response variation using $\ln CVR$ and the meta-regression of $\ln\hat{\sigma}$ with $\ln\overline{x}$ and $Group$

As can be seen in figures 8(A) and (C) there is 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 seens with the pre-intervention scores. 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.

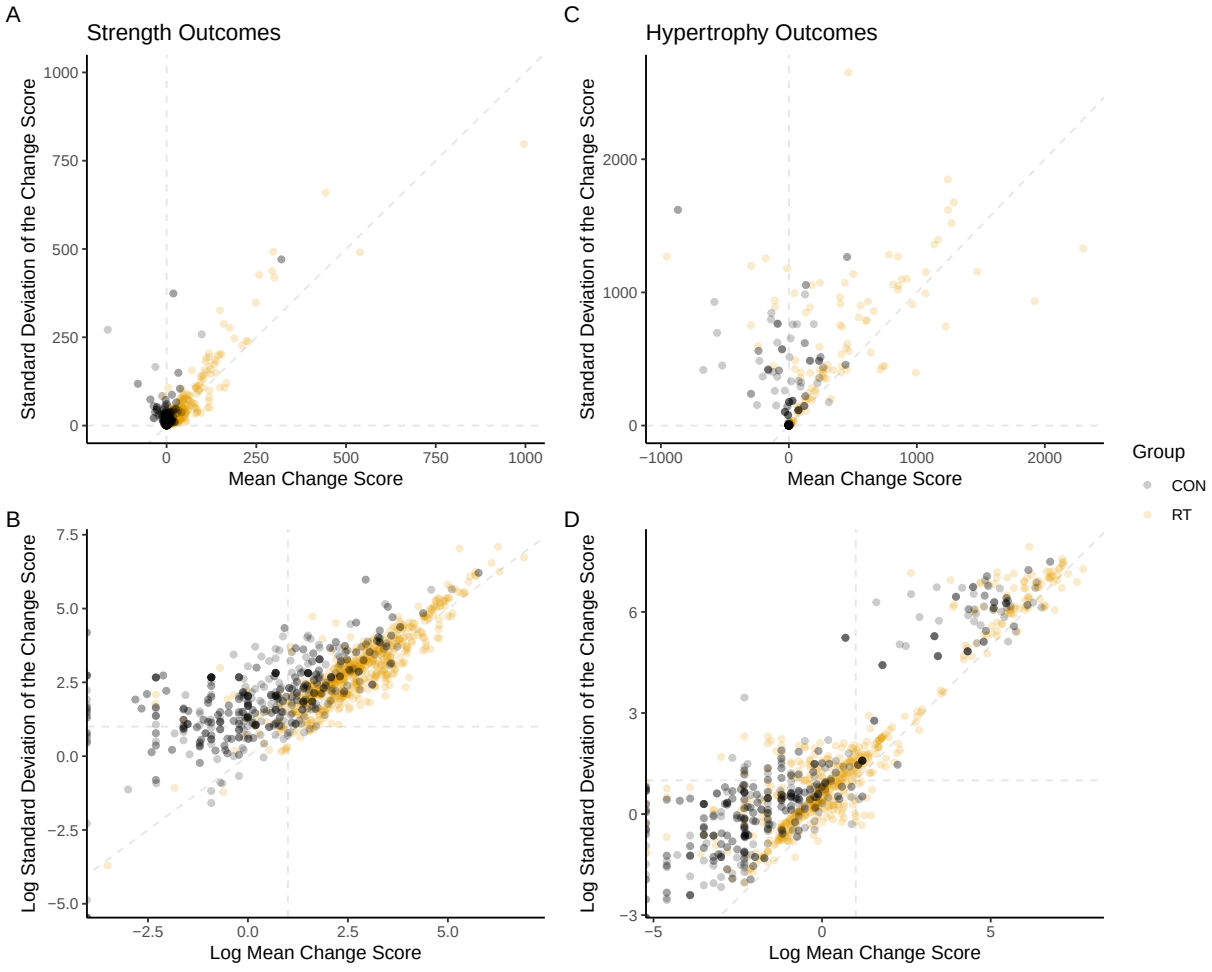

Figure 8: Scatter plots of raw mean and standard deviation of change scores for (A) strength outcomes and (B) hypertrophy outcomes, and of the log mean and log standard deviation of change scores for (C) strength outcomes and (D) hypertrophy outcomes.

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 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 (lnCVR = -0.61 [95%CI: -0.76 to -0.47]; I^2_{study} = 23.18%, I^2_{arm} = 0%, I^2_{effect} = NA%) and hypertrophy (lnCVR = -0.45 [95%CI: -0.61 to -0.29]; I^2_{study} = 10.03%, I^2_{arm} = 0%, I^2_{effect} = NA%) and further there is lower relative heterogeneity between studies in the effect estimates (figure 9).

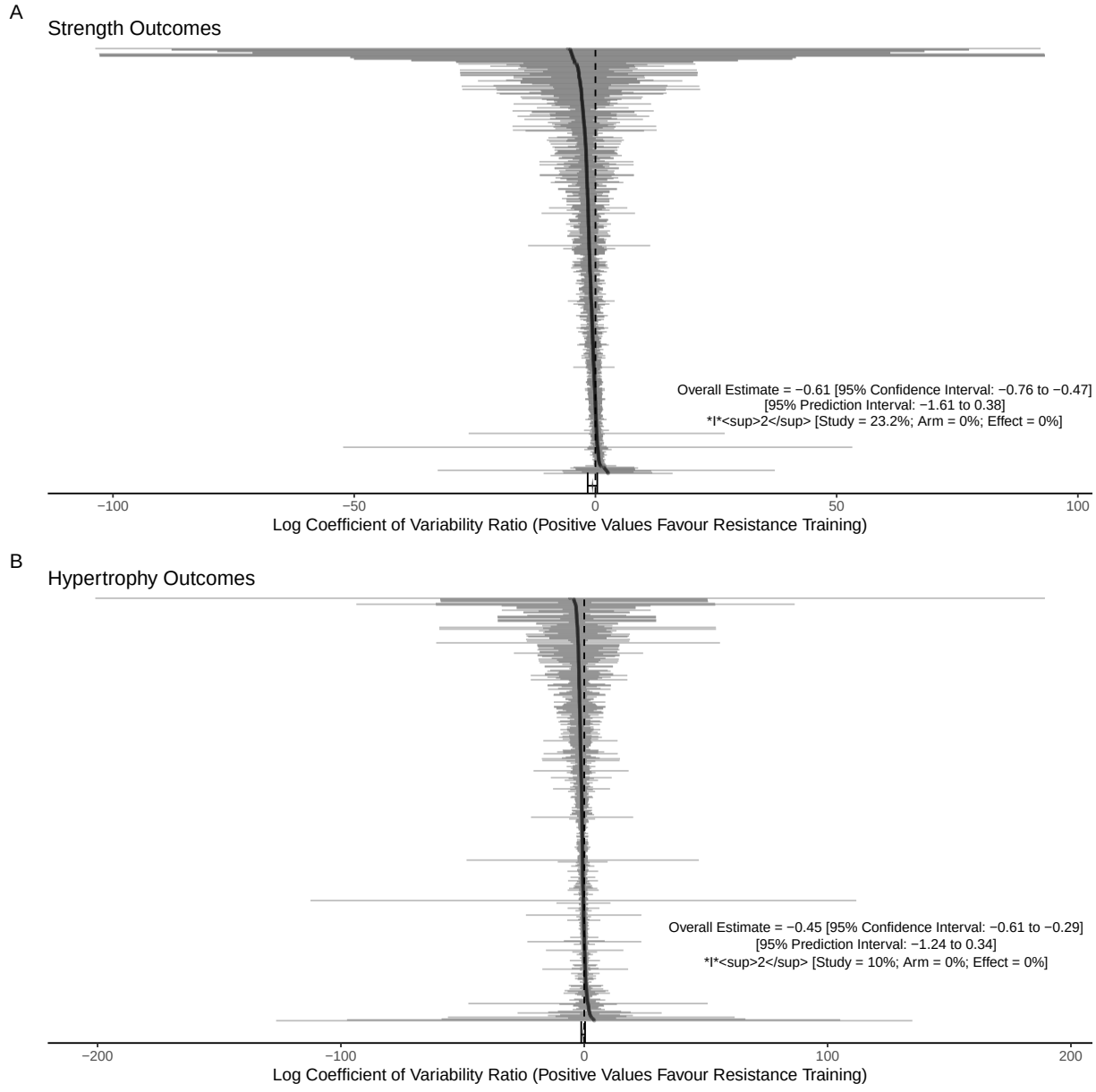

Figure 9: Caterpillar plots of lnCVR effect sizes for strength (A) and hypertrophy (B) outcomes.

There is, however, a potential limitation for the lnCVR 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 lnVR); hence the need to transform the change scores to be positively signed. Secondly, whilst the lnCVR 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 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:

$$\ln CVR_{ijk} = \mu + \tau_i + \tau_j + \tau_k + m_{ij} \tag{18}$$

where $\ln CVR_{ijk}$ is the kth effect size, as in equation (13), for the jth arm ($j = 1, 2, \cdots, N_j$; where N_j is the number of arms) 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, τ_j is the deviation for the jth arm, and τ_k is the deviation for the kth effect which are assumed to be normally 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 model described above (see equation (17) and the paragraph which followed it) to compare the variability in change scores between intervention and control groups using $\ln\hat{\sigma}$ and $\ln\overline{x}$. 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 $Group$.

Given the heteroskedasticity in the change scores means and standard deviations (see figure 8), we fit this model to the dataset[12]. The results were largely similar, albeit slightly attenuated, to those found using the $\ln CVR$ model for strength ($\beta_{\ln\hat{\sigma}[Group \text{ for RT}]} = 0.29$ [95%CI: 0.19 to 0.39]) and hypertrophy ($\beta_{\ln\hat{\sigma}[Group \text{ for RT}]} = 0.18$ [95%CI: 0.09 to 0.26]). See figure 10.

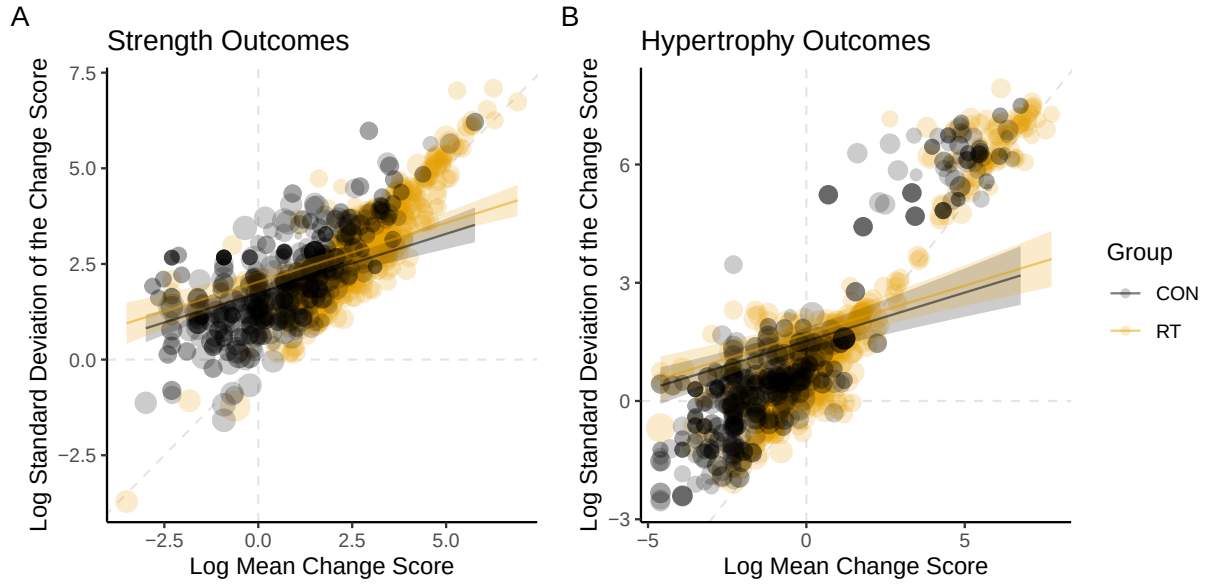

Figure 10: Meta-analytic scatter plot of the log mean and log standard deviation of change scores.

Hopefully it is clear from the regression models here, where we have included both fixed and random predictors as both categorical (i.e., $Outcome$, or $Group$) and continuous (i.e., $\ln\overline{x}$) variables, that there is considerable

[12]Note, as with the models examinine outcome upon baseline scores, we similarly explored $\ln\hat{\sigma}$ with $\ln\overline{x}$ and $Group$ as a predictor with (1) random intercepts for study and arm only, (2) the inclusion of random slopes for $\ln\overline{x}$ by study, and (3) the inclusion of random slopes for $\ln\overline{x}$ 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 the addition of both random slopes improved model fit significantly. Though again we limit presentation in the main text to the simpler model as substantively conclusions were the same.

flexibility in the inclusion of predictors (i.e., meta-regression) when exploring variance through a meta-analytic framework. Models can be fitted to explore not only how study, arm, or effect level characteristics moderate effect size estimates when considering not only effect sizes such as SMDs or lnRR, but also when considering the variance-based effect size statistics and models employed in this article[13].

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 means[14] (**yang_low_2022?**). The examples presented herein used data from RT studies included in a recent meta-analysis published in the *Journal of Sport Sciences* (**polito_moderators_2021?**), which hopefully makes the findings more relatable for researchers in sport and exercise sciences.

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. Using absolute comparisons of variance such as SD_{ir} and lnVR 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 for strength and hypertrophy which is positive, yet combined with the results from SD_{ir} and lnVR we might conclude that while all likely benefit, some benefit more so than others. Indeed, even if for an intervention there was not clear evidence for average intervention effects, exploring variation in their absence might still be important as interventions with large enough variance could imply that the intervention is at least beneficial to some (**usui_meta-analysis_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.

However, similar to the cross-sectional pre-intervention scores reported here (and indeed most physical and biological variables), change scores demonstrated a mean-variance relationship in addition to heteroskedasticity. The likely more appropriate analyses in this case using the lnCVR or meta-regression of ln$\hat{\sigma}$ upon ln$\overline{x}$ 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 interindividual response variation to explain. Interventions, such as RT interventions explored here, 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_meta-analysis_2021?**); that is to say we can assume a constant effect and that the average intervention effect is indicative of the individual intervention effect (**cortes_martinez_constant_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 such recommendations (**steele_higher_2017?**; **steele_long-term_2022?**).

The reason for the apparent reduction in variation after introduction of an RT intervention observed here is not necessarily discernible from this analysis. Perhaps the introduction of an RT intervention

[13]See supplementary materials (https://osf.io/e6vpr) for examples from model estimates for both SMD and lnCVR (used as results for lnCVR and the random slope model were similar) across a range of categorical and continuous predictors for both strength and hypertrophy outcomes. There were no obvious moderators of lnCVR in particular.

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

has indirect effects that reduce other sources of random variance (e.g., diet, other physical activity etc.; (**halliday__resistance__2017?**)), or a ceiling effect on change (i.e., plateau in response; (**steele__long-term__2022?**)) has a constraining effect (**cortes__martinez__constant__2021?**). However, this potentially represents another interesting area of future study regarding 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.

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__individual__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__meta-analysis__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 (**nakagawa__meta-analysis__2015?**; **hopkins__individual__2015?**; **atkinson__true__2015?**; **atkinson__issues__2019?**; **usui__meta-analysis__2021?**; **mills__detecting__2021?**) presented here and echoing the efforts of other recent work (**kelley__are__2022?**; **kelley__are__2020?**; **esteves__individual__2021?**; **bonafiglia__interindividual__2022?**; **steele__slow__2021?**; **fisher__role__2022?**).

6 References