Received: 14 July 2022 | Revised: 14 April 2024 | Accepted: 6 May 2024

DOI: 10.1002/sim.10114

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Standardization and other approaches to meta-analyze differences in means

Will G. Hopkins[1] | **David S. Rowlands**[2]

[1]Professor of Research Design and Statistics (retired), Internet Society for Sport Science, Auckland, New Zealand

[2]Professor of Nutrition, Metabolism, and Exercise, Massey University, Auckland, New Zealand

Correspondence
David S. Rowlands, School of Sport, Exercise, and Nutrition, College of Health, Massey University, Albany Expressway SH17, Albany, Auckland 0632, New Zealand.
Email: d.s.rowlands@massey.ac.nz

Abstract

Meta-analysts often use standardized mean differences (SMD) to combine mean effects from studies in which the dependent variable has been measured with different instruments or scales. In this tutorial we show how the SMD is properly calculated as the difference in means divided by a between-subject reference-group, control-group, or pooled pre-intervention SD, usually free of measurement error. When combining mean effects from controlled trials and crossovers, most meta-analysts have divided by either the pooled SD of change scores, the pooled SD of post-intervention scores, or the pooled SD of pre- and post-intervention scores, resulting in SMDs that are biased and difficult to interpret. The frequent use of such inappropriate standardizing SDs by meta-analysts in three medical journals we surveyed is due to misleading advice in peer-reviewed publications and meta-analysis packages. Even with an appropriate standardizing SD, meta-analysis of SMDs increases heterogeneity artifactually via differences in the standardizing SD between settings. Furthermore, the usual magnitude thresholds for standardized mean effects are not thresholds for clinically important differences. We therefore explain how to use other approaches to combining mean effects of disparate measures: log transformation of factor effects (response ratios) and of percent effects converted to factors; rescaling of psychometrics to percent of maximum range; and rescaling with minimum clinically important differences. In the absence of clinically important differences, we explain how standardization *after* meta-analysis with appropriately transformed or rescaled pre-intervention SDs can be used to assess magnitudes of a meta-analyzed mean effect in different settings.

KEYWORDS

bias, change-score SD, Cochrane, factor effect, meta-analysis, standardized mean difference

1 | INTRODUCTION

Standardization is a method for assessing the magnitude of a difference between two means, devised originally by Jacob Cohen[1(p20)] to compare the means of independent groups. The measure for this comparison is the standardized mean difference:

$$\text{SMD} = \Delta\text{M}/\text{SD}_{\text{Stz}}, \quad (1)$$

where ΔM is the difference in the means, and SD_{Stz} is an appropriate standardizing SD, representing the differences between subjects in either or both groups. The standardized mean differences (SMD) is a dimensionless measure of magnitude capturing the notion that the magnitude of the difference in means should be interpreted according to the overlap of the distributions of the values in each group: large when there is little overlap (because almost every subject in one group has a value greater than that of almost every subject in the other group), and negligible when there is almost complete overlap. The SMD is known as Cohen's *d* after its progenitor, who suggested thresholds that define substantial small, moderate, and large values: 0.20, 0.50, and 0.80, respectively[1(p23)]; other thresholds for an augmented scale have been suggested.[2] Cohen's d is biased high in small samples, owing to underestimation of the population SD by the sample SD. Corrected for small-sample bias, the SMD is sometimes known as Hedges' *g*, when the standardizing SD is the combined SDs of the two groups,[3] or Glass's *g* or *δ*, when the reference or control-group SD is used to standardize.[4]

Gene Glass[5] was the first to suggest the use of SMD to combine mean effects in meta-analysis. When dependent variables in different studies are related by a linear transformation, the numerator and denominator in Equation (1) have the same scaling, which cancels out and allows the disparate effects to be combined. The SMD is also approximately invariant for two variables related by a non-linear transformation, $y = \text{f}(x)$: for sufficiently small ΔM_x evaluated at the mean of x, and for sufficiently small SD_x, $\Delta\text{M}_y \approx \text{f}'(x).\Delta\text{M}_x$, $\text{SD}_y \approx \text{f}'(x)$. SD_x, hence $\Delta\text{M}_y/\text{SD}_y \approx \Delta\text{M}_x/\text{SD}_x$. For example, if $\text{M}_x = 10$, $\Delta\text{M}_x = 0.1\text{M}_x$, $\text{SD}_x = 0.1\text{M}_x$, and $y = x^3$ (the relationship between power and speed with pure fluid resistance to motion), then $\text{SMD}_x = \Delta\text{M}_x/\text{SD}_x = 0.10/0.10 = 1.00$, and from a simple simulation in a spreadsheet (Data S4), $\text{SMD}_y = \Delta\text{M}_y/\text{SD}_y = 334/308 = 1.08$. The SMDs are equal to two decimal places for the smallest important SMD: $\Delta\text{M}_x = 0.02\text{M}_x$, $\text{SMD}_x = \Delta\text{M}_x/\text{SD}_x = 0.20$, and $\text{SMD}_y = \Delta\text{M}_y/\text{SD}_y = 62/308 = 0.20$.

When the meta-analysis is of the difference in the means of a variable that has been measured only once in each subject in two independent groups, only the between-subject SDs of the two groups or their combination are available to compute the SMD. In designs involving changes in the mean of a variable due to an intervention between two measurements in each subject, SDs of post-intervention scores and change scores also become available for standardization. Many authors have recommended or simply assumed that the pre-intervention SD is appropriate,[6-12] but others have stated[12-15] that the SD of change scores is a valid option. Even Cohen was explicit about using the SD of change scores,[16(p539)] but he did not justify this assertion. Surprisingly, in a recent comprehensive investigation of various estimates of the variance of SMDs for meta-analysis,[17] the authors assumed that the two groups had the same population variance, and that the pooling of the sample variances in the two groups was used for standardization, without consideration of whether the variance represented raw scores or change scores; they apparently assumed that mean changes in an intervention should be standardized with the SD of the change scores. In a recent review of the use of SMD in clinical trials, Luo et al.[18] showed large variations in the magnitude of SMDs arising from the different ways means and SDs are combined; these authors did not identify a single best method to calculate the SMD, but noted that "Future methodological studies may dig deeper to see whether it is possible to identify a more universally reasonable approach." The present tutorial provides clinical researchers and meta-analysts with a systematic rationale and guidelines for several such approaches.

First, we argue for the use of a between-subject reference-group, control-group, or pre-intervention SD appropriate to the design of studies contributing mean effects. Next, we provide examples of substantial upward and downward bias arising from use of inappropriate SDs, especially the SD of change scores. We then review the prevalence of the use of different standardizing SDs in meta-analyses published recently in three medical journals. We also review the instructions for meta-analysis of mean effects in the main meta-analysis packages, which have probably contributed to the use of inappropriate standardizing SDs. Finally, we describe several problems with the use of the SMD to combine mean effects in a meta-analysis, even when the standardizing SD is appropriate. We therefore describe alternative approaches: log transformation of factor effects (response ratios) and of percent effects converted to factors; rescaling of psychometrics to percent of maximum range; rescaling with minimum clinically important differences; and standardization *after* meta-analysis with appropriately transformed or rescaled pre-intervention SDs.

2 | APPROPRIATE SDS FOR META-ANALYSIS OF SMDS

2.1 | Cross-sectional studies

When the meta-analysis is of the mean differences between two independent groups in cross-sectional studies, the SDs of the two groups must differ to some extent, so the question arises as to which SD should be used to standardize.

If one group is a reference group (eg, males), then the SMD for another group (eg, females) computed using the reference group's SD will represent the difference between the means of the groups from the perspective of subjects in the reference group: the mean difference relative to the differences they see between themselves. Use of the SD of the other group will result in a different value of the SMD and the alternative perspective. If a single SMD is desired, averaging the two SMDs is one approach, and this SMD represents the difference in means divided by the harmonic mean of the two SDs. Averaging the SDs via weighting of the variances by their degrees of freedom would also be appropriate, but only if there is a reasonable basis for assuming that the difference between the SDs is due only to sampling variation.

Whichever SD is being considered for the SMD in a cross-sectional study, it inevitably includes error of measurement, with contributions from technical error in the measuring instrument and biological variability within subjects. Technical error is random noise arising in the measuring instrument every time it is used to make a measurement, while biological variability is random fluctuations from a subject's underlying true (in the sense of *stable*) value every time the subject is measured. Although the subjects are measured only once in a cross-sectional study, biological variability is nevertheless present in each subject's value.

With the assumption of normality and independence of the true differences between subjects, the within-subject variability, and the technical error, the following model underlies the data:

$$y_{ij} = \mu_i + t_{ij} + s_{ij} + e_{ij}, \tag{2}$$

where y_{ij} is the observed value for the jth subject in group i; $i = 1, 2$ identifies the two groups; $j = 1, 2, \ldots, n_i$ identifies the subjects in the two groups (of sample size n_1 and n_2); μ_i are the population means in each group; t_{ij} is a random effect, distributed $N(0,{\sigma_t}^2{}_i)$, representing true values of the subjects; s_{ij} is a random effect, distributed $N(0,{\sigma_s}^2{}_i)$, representing biological variation within subjects; and e_{ij} is the technical error, distributed $N(0,e^2)$, assumed to have the same variance in each group. If group i is the reference group, the expected value of the observed SD in this group is approximately $\sqrt{({\sigma_t}^2{}_i + {\sigma_s}^2{}_i + e^2)}$. The biological differences between subjects expressed as an SD are given by an estimate either of $\sqrt{({\sigma_t}^2{}_i + {\sigma_s}^2{}_i)}$ or of σ_{ti}, depending on whether the meta-analyst wishes to focus respectively on acute or stable differences between subjects for standardizing. Technical error increases the observed SD above that due to biological differences, so failure to remove technical error results in downward bias of the SMD, and there is also upward bias in heterogeneity, when the technical error differs between studies. The standardizing SD free of technical error is given by:

$$\mathrm{SD_{Stz}} = \sqrt{(\mathrm{SD}^2 - \mathrm{sd}^2)}, \tag{3}$$

where SD is the observed SD and sd is the observed technical error. If the authors of a given study did not provide an estimate of the technical error, the meta-analyst would need to source it from another publication. The technical error is derived from analysis of repeated measurements that do not include biological variability (eg, by analyzing splits of a single biological sample from each subject); it is often presented as a coefficient of variation, in which case it can be converted to the raw units of the SD by dividing by 100 and multiplying by the observed mean in the reference group.

The SE of measurement (SEM) derived from a reliability study, in which subjects are measured on two or more occasions, is an estimate whose expected value is approximately $\sqrt{({\sigma_s}^2{}_i + e^2)}$; hence, if the meta-analyst wishes to use an estimate of the true between-subject SD to standardize, it is given by Equation (3), with sd replaced by the SEM. An alternative formula involves the retest intraclass correlation coefficient (ICC) from a reliability study, with expected value ${\sigma_t}^2{}_i/({\sigma_t}^2{}_i + {\sigma_s}^2{}_i + e^2)$; hence $\mathrm{SD_{Stz}} = \mathrm{SD}\sqrt{(\mathrm{ICC})}$. With an ICC of 0.90, the SMD calculated without removal of the SEM is biased low by a factor of $\sqrt{0.90} = 0.95$ or 5%. See Table 1 for the bias in the SMD with other values of the ICC. The SEM or ICC should come from a reliability study of a sample of subjects drawn from a population like that of the reference group. The time between measurements in the reliability study is an issue. For some measures (eg, blood parameters), measurements that are sufficiently close together will have negligible biological variability; the SEM will therefore represent only technical error, and its removal from the standardizing SD results in an SMD representing the mean difference relative to the acute real differences between subjects. With longer time between measurements, biological variability will contribute to the SEM, and the resulting SMD represents the mean difference relative to stable differences between subjects over the time between the measurements.

TABLE 1 Standardized mean differences (SMD) derived by dividing a mean change (ΔM) by various standard deviations (SD) for various intraclass correlation coefficients (the test–retest correlation in the absence of individual responses), when the true pre-intervention SD is 1.00 unit, ΔM is 1.00 unit, and SD of individual responses is 0.50 units.

									SMD computed using these SD:				
True SD_{Pre}	ICC	SEM	SD_{Pre}	SD_{Δ}	SD_{IR}	SD_{Δ} and SD_{IR}	SD_{Pre} and SD_{Post}	ΔM	True SD_{Pre}	Observed SD_{Pre}	SD_{Δ}	SD_{Δ} and SD_{IR}	SD_{Pre} and SD_{Post}
1.00	0.99	0.10	1.01	0.14	0.50	0.52	1.07	1.00	1.00	0.99	7.04	1.92	0.94
1.00	0.95	0.23	1.03	0.32	0.50	0.60	1.09	1.00	1.00	0.97	3.08	1.68	0.92
1.00	0.90	0.33	1.05	0.47	0.50	0.69	1.11	1.00	1.00	0.95	2.12	1.46	0.90
1.00	0.70	0.65	1.20	0.93	0.50	1.05	1.25	1.00	1.00	0.84	1.08	0.95	0.80
1.00	0.50	1.00	1.41	1.41	0.50	1.50	1.46	1.00	1.00	0.71	0.71	0.67	0.69
1.00	0.30	1.53	1.83	2.16	0.50	2.22	1.86	1.00	1.00	0.55	0.46	0.45	0.54

Note: ICC, intraclass correlation coefficient: $TrueSD_{Pre}^2/SD_{Pre}^2$. Magnitude scale for SMD: <0.20, trivial, 0.20-0.59, small; 0.60-1.19, moderate; 1.20-1.99 large; 2.00-3.99, very large; ≥4.00, extremely large.[2] SD_{IR}, SD of individual responses: $\sqrt{(SD_{Post}^2 - SD_{Pre}^2)}$. SD_{Post}, observed post-intervention SD: $\sqrt{(SD_{Pre}^2 + SD_{IR}^2)}$. SD_{Pre} and SD_{Post}, mean of observed pre- and post-intervention SD: $\sqrt{[(SD_{Pre}^2 + SD_{Post}^2)/2]}$. SD_{Pre}, observed pre-intervention SD: $\sqrt{(TrueSD_{Pre}^2 + SEM^2)}$. SD_{Δ} and SD_{IR}, SD of change scores including individual responses: $\sqrt{(SEM^2 + SEM^2 + SD_{IR}^2)}$. SD_{Δ}, SD of change scores without individual responses: $\sqrt{(SEM^2 + SEM^2)}$. SEM, standard error of measurement: $TrueSD_{Pre}\sqrt{[(1–ICC)/ICC]}$. $TrueSD_{Pre}$, pre-intervention SD free of measurement error: $\sqrt{(SD_{Pre}^2 - SEM^2)}$.

2.2 | Post-only controlled trials

Some controlled trials are analyzed in a manner like that of a cross-sectional study: the effect of an intervention is calculated as the difference in the means of a dependent variable measured only once in the control and experimental groups, after the intervention. This kind of post-only analysis provides better precision for the mean effect than the more usual analysis of the difference in the mean changes between pre and post measurements, when the error of measurement over the time-frame of the intervention is greater than the error-free between-subject SD (eg, section 6.5.2.8 in the Cochrane Handbook[14]).

The following model is assumed to underlie the data:

$$y_{ij} = \mu_i + t_{ij} + s_{ij} + r_{ij} + e_{ij}, \tag{4}$$

where $i = C, E$ now refers to control and experimental groups; y_{ij}, μ_i, t_{ij}, s_{ij}, and e_{ij} are defined as for Equation (2); and r_{ij} is a random effect, distributed $N(0,{\sigma_r}^2{}_i)$, representing individual responses to the control and intervention treatments. The observed variances in the groups have expected values ${\sigma_t}^2{}_i + {\sigma_s}^2{}_i + {\sigma_r}^2{}_i + e^2$. As in a cross-sectional study, the SD of the control and experimental groups in a post-only controlled trial are both available for computing the SMD. An individual in the population under study would or should be interested in the extent to which an intervention changes their value relative to the values of untreated individuals. The SD of the control group provides an estimate of this SD, with the reasonable assumption that individual responses to the control (${\sigma_r}^2{}_C$) are negligible. Use of this SD is also justified for clinically important measures: to the extent that the differences between individuals are associated with health outcomes, it is worth doing an intervention to change the measure, and the bigger the mean change relative to the differences, the more clinically important is the intervention. The SMD standardized with the control SD captures this importance. Again, the contribution of error of measurement to the standardizing SD (arising only from technical error or including changes in individuals that would occur in the control group over the duration of the intervention or some other duration) can be removed from the standardizing SD, as described in Section 2.1, Equation (3).

The SD in the experimental group is likely to be different from that in the control group for reasons other than sampling variation: it is usually larger, owing to individual responses to the intervention (σ_{rE}) adding to the between-subject SD, but it can be smaller (represented by a negative estimate for ${\sigma_r}^2{}_E$ when negative variance is allowed in mixed models), if the intervention somehow compresses the range of individual values, for example by raising them towards an upper limit or reducing them towards a lower limit. Use of the SD of the experimental group would therefore bias the SMD, and it would also introduce artifactual heterogeneity, when individual responses or compression of responses differ between studies.

Pooling of the SD of both groups cannot be justified, even if their sample values seem similar, since the intervention cannot be assumed to have negligible effect on individual responses or compression of responses.

Individual responses to the intervention should be of interest, but their magnitude should be assessed separately as a SD, SD_{IR}. The expected values of the observed SD^2 in the two groups are ${\sigma_t}^2{}_C + {\sigma_s}^2{}_C + {\sigma_r}^2{}_C + e^2$, and ${\sigma_t}^2{}_E + {\sigma_s}^2{}_E + {\sigma_r}^2{}_E + e^2$. Assuming that the subjects are well randomized or assigned in a balanced fashion to the two groups, then ${\sigma_t}^2{}_C = {\sigma_t}^2{}_E$ and ${\sigma_s}^2{}_C = {\sigma_s}^2{}_E$, so the expected value of the observed experimental minus observed control variance is ${\sigma_r}^2{}_E$ - ${\sigma_r}^2{}_C$. Hence an estimate of the SD of individual responses, SD_{IR}, is given by:

$$SD_{IR} = \sqrt{(SD_E^2 - SD_C^2)}, \tag{5}$$

where SD_E and SD_C are the observed standard deviations in the experimental and control groups. Note that this SD_{IR} represents individual responses additional to those arising from a control intervention; replacement of SD_C with a between-subject SD measured without a control intervention in individuals randomly chosen from the same population would obviate this problem. The SD_{IR} is free of measurement error and represents typical individual differences from the mean effect of the intervention; the magnitude of this SD can also be assessed by standardization (by dividing it by the same SD used to standardize the mean effect), but the magnitude thresholds are half those of the SMD.[19] Separate meta-analysis of the SD_{IR} is possible but beyond the scope of this tutorial.

2.3 | Time series

There are four designs in which SDs of change scores are available for standardization prior to meta-analysis. In the simplest of these designs, a time series, only one group of subjects is measured pre and post intervention. Only one kind of change score is therefore available. The following model underlies the data:

$$y_{ij} = \mu_i + t_j + s_{ij} + r_{ij} + e_{ij}, \tag{6}$$

where $i = 1, 2$ identifies pre- and post-intervention measurements; μ_i are the pre and post means; t_j is a random effect, distributed $N(0,{\sigma_t}^2)$, representing true (stable) differences between subjects; s_{ij} is a random effect, distributed $N(0,{\sigma_s}^2)$, representing short-term variation within subjects (with the same variance pre and post intervention); r_{ij} is a random effect, zero when $i = 1$ and distributed $N(0,{\sigma_r}^2)$ when $i = 2$, representing individual responses to the intervention; and e_{ij} is the technical error, distributed $N(0,e^2)$. The variance of the change scores has expected value $2{\sigma_s}^2 + {\sigma_r}^2 + 2e^2$. The observed variance ($SD_\Delta{}^2$) therefore has contributions from individual responses in the post measurement ($SD_{IR}{}^2$, expected value ${\sigma_r}^2$) and SE of measurement in pre and post measurements ($2SEM^2$, expected value $2{\sigma_s}^2 + 2e^2$); hence SD_Δ is given by:

$$SD_\Delta = \sqrt{(SD_{IR}^2 + 2SEM^2)} \tag{7}$$

By the same argument as for the post-only controlled trial, the only SD that should be used for standardizing the mean change is the pre-intervention SD (SD_{Pre}), free of technical and possibly measurement error estimated as described in Section 2.1, Equation (3). Use of the post-intervention SD (SD_{Post}), or a pooled SD_{Pre} and SD_{Post}, produces an SMD contaminated by the SD_{IR}. Use of the SD_Δ produces an SMD scaled according to the combination of individual responses and measurement error, which cannot represent the clinical importance or any other real-world interpretation of the effect of an intervention, and it can grossly overestimate the true effect magnitude (Section 3 below; see also References 10,12). Removal of measurement error from the SD_Δ (via Equation (7) and a separately sourced SEM) scales the SMD purely with the SD_{IR}, given by:

$$SD_{IR} = \sqrt{(SD_\Delta^2 - 2SEM^2)}. \tag{8}$$

But the result is also an unrealistic representation of the magnitude of the effect; for example, if the individual responses were negligible compared with the mean change, the SMD would be huge. Individual differences in the effect of an intervention are of course important, but they should be dealt with in a separate analysis, not by making them the denominator for standardizing the mean change. Furthermore, if the SD_{IR} were used for standardizing the change in the

mean, it would not make sense to use the SD_{IR} to standardize itself, nor would it make sense to use the SD_{IR} to standardize the mean change and some other SD to standardize the SD_{IR}.

2.4 | Post-only crossovers

In this design, subjects are divided into two groups, which are measured after receiving either a control or experimental intervention. After a sufficient period of time to wash out the effects of the intervention, the groups receive the other intervention and are then measured again. The underlying model is similar to that of a time series:

$$y_{ij} = \mu_i + t_j + s_{ij} + r_{ij} + e_{ij}, \tag{9}$$

except that $i = \mathrm{C}, \mathrm{E}$ now identifies control and intervention treatments, and r_{ij} is a random effect, distributed $N(0,{\sigma_r}^2{}_i)$, representing individual responses to the control and intervention treatments. An SD of changes between control and the intervention treatments is available, but with variance $2{\sigma_s}^2 + {\sigma_r}^2{}_C + {\sigma_r}^2{}_E + 2e^2$, use of this SD for standardizing the change in the mean between control and interventions would have no clinically or practically meaningful interpretation. As in a post-only controlled trial, the SD following the control condition is the only justifiable SD for standardizing, and the issues of contributions to this SD of individual responses to the control and of error of measurement are the same. If washout of the intervention is not complete and differs between individuals in the group receiving the intervention first, the SD following the control condition will be inflated by the variable amount of the effect of the intervention not washed out, in which case only the control-condition SD of the group receiving the control condition first should be used for standardizing.

2.5 | Pre-post controlled trials

In this most common of the repeated-measures designs, subjects are typically randomized or assigned in balanced fashion to control and intervention groups, which are measured before and after the control and intervention treatments. The model is similar to that for post-only controlled trials, with the addition of a subscript $k = 1, 2$ to denote pre- and post-intervention measurements:

$$y_{ijk} = \mu_{ik} + t_{ij} + s_{ijk} + r_{ij} + e_{ijk}. \tag{10}$$

There are now two change-score SDs, one from each group, with variances the same as those in a time series: $2{\sigma_s}^2 + {\sigma_r}^2{}_i + 2e^2$. Neither of these SDs, nor their combination, makes any sense for standardizing, for reasons already discussed. The most appropriate SD is SD_{Pre}: the SD of all the subjects in the pre-test, or if that is not available, the mean of the pre-test SDs in the two groups derived by weighting their variances by their degrees of freedom. SD_{Pre} has variance ${\sigma_t}^2 + {\sigma_s}^2 + e^2$, and the variance of the change scores in the control group, ${SD_\Delta}^2{}_C$, provides an estimate of $2({\sigma_s}^2 + e^2)$, assuming negligible individual responses to the control condition. But ${\sigma_s}^2 + e^2$ is the expected value of the square of the SE of measurement, SEM^2. Hence, if the true value of SD_{Pre} (ie, free of measurement error) is used to standardize, it is given by:

$$SD_{Stz} = \sqrt{\left(SD_{Pre}^2 - SEM^2\right)} = \sqrt{\left(SD_{Pre}^2 - SD_{\Delta C}^2/2\right)}. \tag{11}$$

The difference in the variances of the change scores has expected value $2{\sigma_s}^2 + {\sigma_r}^2{}_E + 2e^2 - (2{\sigma_s}^2 + {\sigma_r}^2{}_C + 2e^2) = {\sigma_r}^2{}_E$, assuming ${\sigma_r}^2{}_C$ is negligible; hence, assessment of individual responses is possible via a SD free of error of measurement, given by:

$$SD_{IR} = \sqrt{\left(SD_{\Delta E}^2 - SD_{\Delta C}^2\right)}, \tag{12}$$

where $SD_{\Delta E}$ and $SD_{\Delta C}$ are the standard deviations of change scores in the experimental and control groups.[19]

2.6 | Pre-post crossovers

The final design providing change scores that could be used for standardization is the pre-post crossover, in which subjects are divided into two groups measured before and after either an interventional or control condition. After a sufficient period of time to wash out the effects of the intervention, the groups are measured before and after receiving the other condition. The model is the same as Equation (10), except that the true values of the subjects are defined by t_j, since the same subjects experience both conditions. There are now posted minus pre change scores for the intervention and control, intervention minus control change scores in the pre-test, and intervention minus control change scores in the post-test. The SDs of these change scores have different contributions of measurement error and individual responses, and as such, none is suitable alone or in combination for standardization. The most appropriate SD is that of all the subjects in their first pre-test, with removal of technical and measurement error estimated from the pooled SD of change scores with the control condition, as in pre-post controlled trials. Individual responses can be estimated with the same formula as in the pre-post controlled trial by pooling the SDs of post-pre change scores of the intervention in both groups and those of the control condition in both groups.

2.7 | SMDs for meta-analysis of mixed designs

A further argument can be adduced for not using SDs of change scores (SD_Δ) to standardize, when the meta-analysis includes post-only controlled trials along with studies involving repeated measurement. The biological effect of the intervention on individuals is the same in the different designs, and all the effects should therefore be included in the meta-analysis, but obviously, an SMD computed with the SD of the control group in a post-only controlled trial is not commensurate with SMDs computed with change scores in the other designs. Furthermore, the SMDs in the other designs will differ, depending on the contributions of error of measurement and individual responses to whatever single or composite standardizing SD the meta-analyst chooses in the different designs. The SMD in the post-only controlled trial could be made commensurate by converting it to an SMD using an SD_Δ based on an assumption or imputation of a re-test correlation or of error of measurement and individual responses, but the use of control or pre-intervention SDs across all the studies, preferably after removal of technical error and within-subject error of measurement, makes the SMDs commensurate and realistic.

3 | BIAS IN SMDS WITH INAPPROPRIATE SDS

In Table 1 we provide examples of bias in the SMD arising from use of various standardizing SDs. For simplicity, the examples are for changes in a single group resulting from an intervention (a time-series design), and we subsequently discuss the bias in controlled trials.

We have calculated the bias relative to the SMD calculated with what we have argued to be the correct standardizing SD ($TrueSD_{Pre}$): the between-subject SD of pre-intervention scores free of measurement error arising from instrumentation and short-term biological variation of no functional significance. In our experience, it would be unusual for a measure to be in common use, if such error resulted in a reliability or retest ICC less than ~0.30, the lower limit of ICC in Table 1. On the other hand, very high reliabilities are not uncommon, hence the highest ICC of 0.99. If $TrueSD_{Pre}$ is 1.00 unit, and the mean change is 1.00 unit, the resulting SMD is 1.00, a moderate effect using a modified and augmented version[2] of Cohen's scale.[16(p25,26)] For the highest reliability, SMDs using the observed pre-intervention SD or pooled pre- and post-intervention SDs show only slight downward bias, but SMDs using the SD_Δ are biased high to an extremely large value, or to a large value when individual responses with an SD of half the true pre-intervention SD are included in the change scores. At the lowest reliability, the SMDs are all biased low to small values, when SDs other than the true SD are used to standardize.

As stated previously, the SD_Δ in a controlled trial includes technical error, individual biological changes that occur naturally in the time between pre and post measurements, and individual responses to the intervention (including any placebo or active treatment in the control group). The bias in the SMD using a pooled SD_Δ of control and experimental groups will therefore be less than that shown in Table 1 for no individual responses, but the bias could be similar to, or greater than, that shown in Table 1 for change scores with individual responses. Meta-analysts who pool the post-intervention SDs in experimental and control groups to derive the SE of the mean effect, and who then use the pooled

SDs to standardize, will incur bias like that shown in Table 1 for combining the pre- and post-intervention SDs, since the pre-intervention SD will be similar to the post-intervention SD in the control group.

4 | PREVALENCE OF INAPPROPRIATE SMDS IN META-ANALYSES

We have long been aware that some meta-analysts used the SD of change scores to standardize mean effects, but we had only vague notions of the prevalence of this misuse and the extent of bias it can produce. After noticing the error in a recent meta-analysis by Nunes et al.[20] in the *Journal of Cachexia, Sarcopenia, and Muscle* (*JCSM*), we examined the study,[21] in that meta-analysis with the apparently the largest SMD, 2.43 (a very large effect). When we calculated the SMD with the pooled pre-intervention SDs, the SMD reduced to 0.28 (a small effect). This astonishing difference between the SMDs motivated us to further investigate the prevalence of the use of change-score and other SDs to calculate SMDs in recent meta-analyses. However, a reviewer of our tutorial article pointed out that the authors of this study[21] calculated the SD of post-pre change scores incorrectly as the post-pre change in the SDs. We subsequently chose the study[22] in the Nunes et al. meta-analysis with the next largest SMD, 1.03 (a moderate effect) to exemplify the calculations. We arrived at this same value using the means and SDs of change scores of lean body mass in a placebo group (0.70 ± 1.14 kg, $n = 13$) and collagen supplementation group (2.56 ± 2.22 kg, $n = 12$). The difference in the mean changes is 1.86 kg, the pooled change-score SD is 1.74 kg, so the SMD is $1.86/1.74 = 1.07$, reducing to 1.03 after multiplying by the bias-correction factor of 0.97. The pre-intervention SDs in placebo and collagen groups were 4.0 and 5.7 kg, giving a pooled pre-intervention SD of 4.9 kg. The magnitude of the change-score SD in the placebo group (1.14) is much smaller than this pooled SD, even though it includes individual-responses to a resistance-training co-intervention, showing that the measure of lean mass is highly reliable over the period of the intervention: the error of measurement is at most only $1.14/\sqrt{2} = 0.81$ kg, and the ICC is $(4.9^2\text{-}0.81^2)/4.9^2 = 0.973$. The error-free pre-intervention SD is $\sqrt{(4.9^2 - 0.81^2)} = 4.83$ kg, which is only slightly smaller than the observed SD. The SMD recalculated using the pooled pre-intervention SD is 0.38, increasing to 0.39 using the error-free pre-intervention SD, and both reduce to 0.37 (a small effect) after correction for small-sample bias. The SMD calculated correctly with pooled pre-intervention SDs is thus only about one-third of the value calculated with change-score SDs.

Our investigation comprised a Pubmed search of all articles containing at least one meta-analysis in *JCSM* between its inception in 2016 and June 2022 (21 articles) and the most recent 30 articles (to June 2022) containing at least one meta-analysis in a pure and in an applied medical journal: *Journal of American Medical Association* (*JAMA*) and *Sports Medicine*. The results are summarized in Table 2 and detailed in Table S1.

In a total of 116 meta-analyses across the 81 articles, about one-third ($41/116 = 35\%$) involved standardization as a method of analysis of differences of means. In those using standardization, the pooled SD of change scores was most popular (around two-fifths) and pooled post-intervention SDs accounted for around a quarter of the other analyses. Around one-tenth of analyses each used either the pooled pre- and post-intervention SD or the pooled pre-intervention SD. Only around two-tenths of all meta-analysis involved the analysis of raw means. There were no major differences between the three journals in most of these proportions, but *JAMA* articles used standardization least frequently ($9/51 = 18\%$). In relation to alternative approaches to standardization described in Section 7 below, there were only two meta-analyses of percent or factor mean differences, two meta-analyses of rescaled psychometrics, and two meta-analyses in which standardization was used after meta-analysis to assess effect magnitude.

A disturbing feature of the majority of meta-analyses we reviewed was the lack of any description of, or citation for, the approach to calculate the SMD. We had to resolve the approach either through in-depth investigation of the meta-analysis datasets, back calculation, reference back to original cited study data, or by communication with the corresponding authors via email, some of whom replied that they did not know the method of calculation and relied upon the outputs of software. Apparently, no authors considered excluding technical error or SE of measurement from the estimate of the standardizing SD.

5 | SMDS IN POPULAR META-ANALYSIS PACKAGES

The popular software packages for meta-analysis have exacerbated errors in the use of standardization by allowing for the SD of change scores to be used to standardize and by allowing for pooling of SDs of experimental and control groups. R's metafor has an option for pre-intervention SDs to standardize while retaining the precision inherent in deriving the SE

TABLE 2 Numbers (n) and proportions of meta-analyses of various effects either not using or using standardization in recent issues of three medical journals.

	Analysis not using standardization				Analysis using standardization					
	Raw mean differences	Psycho-metric mean differences[a]	Percent or factor mean differences	Other effects[b]	Pooled SD of change scores	Pooled post intervention SD	Pooled pre and post intervention SD	Pooled pre intervention SD	Standardization after meta-analysis	Method not certain
JCSM ($n = 28$)	3	0	1	15	6	2	0	0	0	1
JAMA ($n = 51$)	10	2	0	28	3	3	0	3	1	1
Sports Med ($n = 37$)	8	0	1	7	7	6	4	2	1	1
Total meta-analyses ($n = 116$)	21	2	2	50	16	11	4	5	2	3
Proportion of meta-analyses (%)	18	2	2	43	14	9	3	4	2	3
Proportion (%) of analyses using standardization ($n = 41$)					39	27	10	12	5	7

Abbreviations: JAMA, Journal of the American Medical Association; JCSM, Journal of Cachexia, Sarcopenia, and Muscle; Sports Med, Sports Medicine.

Note: See Data S1 for details and a full reference list of articles that comprise the sample.

[a]Scaled or rescaled to 0-10 or 0-100.

[b]Odds, risks, proportions, hazards, counts, correlations, SDs, or their ratios or differences.

from the SD_Δ.[23(p58)] All packages allow for meta-analysis of SMDs following insertion of SMDs and their SEs calculated by the user, so we finish this section by providing equations for the SE of the SMD and its degrees of freedom with any standardizing SD.

The most popular software for meta-analysis in our survey (38%, Table S1) was R, with the metafor package the most popular.[24] The documentation for metafor[23] provides a function called *escalc* (p. 72) for calculation of SMDs, which has two sets of options. In the first set (p. 76), the user inputs a single mean and SD (either raw scores or change scores) for each of the two groups in each study. The difference in the means is calculated either as a raw difference, as a ratio via log transformation, or as a standardized mean difference using the SD of one group or the SD of both groups combined in a choice of several ways. Thus, if raw scores from an intervention are inputted, the means and SD have to be the post-intervention values of the experimental and control groups; in this case the SD of the control group can be chosen to standardize (using measure = "SMD1"), but change scores cannot be used to estimate the SE, and precision is therefore compromised for most dependent variables. If means and SD of change scores are inputted, precision is not compromised, but the SMD is computed with the SD of change scores. In the second set of options (p. 84), each study consists of a single group with pre and post means and SD (ie, an uncontrolled time series). The user also inserts a separately derived or assumed value for the retest correlation. The user can then choose to standardize using either the pre-intervention SD (using measure = "SMCR"), an average of pre and post SDs, or the change-score SD (computed via the correlation). This set of options therefore allows for the correct pre-intervention SD to be used for the SMD and for better precision based on change scores (although the user may have to derive the retest correlation from the change-score SD and pre- and post-intervention SDs). Alternatively, the user inserts the mean and SD of change scores as the pre-intervention values, inserts zeros for the post-intervention values, and inserts zero for the correlation, but in this case, standardization is performed only with the SD of change scores. With controlled trials, the data for intervention and a control groups would have to be inserted separately, and the difference between the mean changes would be estimated by including an intervention fixed effect in a meta-regression; the repeated measurement in each study should be accounted for with an additional random effect using the rma.mv function.

Stata was the second most popular software for meta-analysis (24%). Stata provides several options for standardizing differences in mean changes between experimental and control groups,[25(p88)] including use of the SD of the control group. However, no mention is made of whether the means and SD input by the user refer to raw scores or change scores, so the computed SMDs correspond to the first set of options in metafor, described above. There is no option for use of pre-intervention SD to standardize and for change-score SD to compute SEs.

The RevMan software[26] was used next frequently (16%). The accompanying Cochrane handbook for systematic reviews of interventions is a widely cited source of meta-analytic methods. It includes a section on standardization (6.5.1.2), in which the expressions "between-participant variability in outcome measurements" and "the standard deviation of outcome among participants" are used to refer to the standardizing SD.[14] These expressions could represent either the SD of the post-intervention scores or the SD of change scores, but the ambiguity is not resolved. To meta-analyze differences in mean changes in controlled trials, the means and SD of change scores must be inserted into RevMan, so there is an extensive section (6.5.2.8) on derivation of the SD of change scores, for use when authors of published studies do not provide it. In a later section (10.5.1) it is stated unambiguously that "the[se] SDs are used to standardize the mean differences to a single scale." Yet confusingly, in section 6.2.5.2 there is the following statement, which appears to indicate that standardizing should be done with a between-subject SD, not a change-score SD:

> To overcome problems associated with estimating SDs within small studies, and with real differences across studies in between-person variability, it may sometimes be desirable to standardize using an external estimate of SD. External estimates might be derived, for example, from a cross-sectional analysis of many individuals assessed using the same continuous outcome measure (the sample of individuals might be derived from a large cohort study).

Only about one-tenth (9%) of meta-analysts used the Comprehensive Meta-Analysis package (CMA). The documentation on standardization is sparse.[27(p70)] Users who wish to combine studies using standardization apparently have to input the sample size in each group and the SMD they have calculated themselves from the means in the two groups and the SDs. The CMA program then derives the SE from the SMD and the sample sizes, so it produces a meta-analyzed SMD similar to the first set of options in metafor, standardized with post-intervention SDs or with SDs of change scores, depending on how the user performed the standardization.

The Statistical Analysis System (SAS), the Statistical Package for the Social Sciences (SPSS) and Excel were the other software packages used by the meta-analysts (4%, 3%, and 1%, respectively), with a further 4% indeterminate. We recommend use of the general linear mixed model (Proc Mixed) in SAS: there is an elegant approach in which the residual variance is held to unity[28] (eg, Rowlands et al.,[29] Patten et al.[30]), and the unbounded random-effect variances come with SEs that allow realistic and thorough assessment of heterogeneity.

Meta-analysts who wish to use standardization to combine mean effects, despite the problems we have identified in Section 6 and our advice in Section 7, can avoid any doubt about the way the package handles standardization and can use any standardizing SD by expressing all the effects as SMD calculated with an appropriate between-subject SD before inserting them into the package. The SEs of each study SMD can be estimated via the following equation, which was adapted from Equation (22) in Lin and Aloe[17] to allow different SDs in the intervention and control groups (and which can be derived as a first-order approximation by differentiating Equation (1)):

$$\mathrm{SE}^2 = \left[(\mathrm{SD}_1^2/\mathrm{n}_1 + \mathrm{SD}_2^2/\mathrm{n}_2/\mathrm{SD}_{\mathrm{Stz}}^2\right] + \left[\mathrm{SMD}^2/(2\mathrm{n}_{\mathrm{Stz}})\right]. \quad (13)$$

In this equation, the first bracketed term is the contribution to the standard error due to the uncertainty in the difference in means; SD_1 and SD_2 are the standard deviations in the two groups, n_1 and n_2 are the sample sizes in the two groups, and $\mathrm{SD}_{\mathrm{Stz}}$ is the appropriate standardizing SD (as described in Section 2 above, or it can come from another study of a similar population). The second bracketed term is the contribution due to the uncertainty in $\mathrm{SD}_{\mathrm{Stz}}$; $\mathrm{n}_{\mathrm{Stz}}$ is its sample size. In a cross-sectional study or post-only controlled trial, SD_1 and SD_2 are raw SDs; in a timeseries or post-only crossover, SD_1 is the SD of change scores, n_1 is the number of change scores, and SD_2 and n_2 are omitted; in a pre-post parallel-groups controlled trial or crossover, SD_1 and SD_2 are the standard deviations of change scores in the intervention and control groups or treatments. If these change scores are not provided in a study, they can be derived easily from confidence intervals, *t* statistics, or exact *P*-values; if the only inferential information for the mean change is a *P*-value inequality, we recommend imputing these SD via the SE of measurement in similar studies, using a panel of cells in a spreadsheet for estimating sample size.[31] The SMD is biased high in small samples, owing to downward bias in the standardizing SD; the bias should be removed by multiplying the SMD and its SE by this factor: $1\text{-}3/(4\mathrm{DF}_{\mathrm{Stz}}-1)$, where $\mathrm{DF}_{\mathrm{Stz}}$ is the degrees of freedom of the standardizing SD.[3]

The degrees of freedom of SE is not required for meta-analysis, but it is required for estimation and presentation of the confidence limits of each SMD; the degrees of freedom can be calculated by applying the well-known Satterthwaite approximation[32] to the two bracketed terms comprising the SE in Equation (13):

$$\mathrm{DF}_{\mathrm{SE}} = \mathrm{SE}^4/\left((\text{first term})^2/\mathrm{DF}_{\Delta} + (\text{second term})^2/\mathrm{DF}_{\mathrm{Stz}}\right). \quad (14)$$

In this equation, DF_{Δ} is the degrees of freedom of the mean effect, itself given by the Satterthwaite approximation when the SDs of the two groups are assumed to be different (as they should be).

6 | OTHER PROBLEMS WITH THE SMD FOR META-ANALYSIS

Even when scales or measuring instruments and units are the same in different studies, the standardizing SD will differ between samples of different populations and between differently biased samples of the same population. Use of the SMD to combine study-estimates for meta-analysis will therefore contribute artifactually to heterogeneity of effects that are otherwise biologically similar. Differences in $\mathrm{SD}_{\mathrm{Stz}}$ arising purely from sampling variation will contribute to observed differences in the SMD between studies, but the increase in their SEs from the second term in Equation (13) will account for such differences, and the resulting estimate of true heterogeneity will not be biased upwards. Real differences in the $\mathrm{SD}_{\mathrm{Stz}}$ will contribute non-biological artifactual heterogeneity, the magnitude of which will depend not only on the real differences but also the meta-analyzed mean SMD and the extent of true heterogeneity. If the true differences in the standardizing SD ($\mathrm{SD}_{\mathrm{Stz}}$) between studies is a factor SD of f (ie, the SDs differ from the mean $\mathrm{SD}_{\mathrm{Stz}}$ by typically ×/÷f), and the mean effect is ΔM, then the SMD for studies with an $\mathrm{SD}_{\mathrm{Stz}}$ above or below the mean $\mathrm{SD}_{\mathrm{Stz}}$ by the factor f is $\Delta\mathrm{M}/(\mathrm{SD}_{\mathrm{Stz}}\times/\div\mathrm{f}) = \mathrm{SMD}\div/\times\mathrm{f}$. If SMD = 0, there is obviously no artifactual heterogeneity, irrespective of f. If f = 1.20 and mean SMD = 1.00 (moderate), the SMD for studies with typically smaller $\mathrm{SD}_{\mathrm{Stz}}$ will contribute an SMD of $1.00 \times 1.20 = 1.20$ to the meta-analysis, while studies with typically larger $\mathrm{SD}_{\mathrm{Stz}}$ will contribute an SMD of $1.00 \div 1.20 = 0.83$. This range in SMD defined by the factor SD can be represented approximately as 1.00 ± 0.19 (mean ± SD). Thus, in the absence of

any true heterogeneity, the heterogeneity provided by the meta-analysis is entirely artifactual, and the value of 0.19 is small in magnitude (using magnitude thresholds that are half those of the SMD[19]). In the presence of true heterogeneity in SMD, the contribution of artifactual heterogeneity to the observed meta-analyzed heterogeneity diminishes; for example, if the true heterogeneity is a SD of 0.45 (mid-range moderate), the observed heterogeneity in the meta-analysis is $\sqrt{(0.45^2 + 0.19^2)} = 0.49$, a practically negligible increase. A larger meta-analyzed mean SMD and/or larger between-study differences in the standardizing SD would result in a larger artifactual contribution to the heterogeneity.

Use of the SMD also increases the uncertainty in the mean effect arising from uncertainty in the standardizing SD (the second bracketed term in Equation (13)). This uncertainty can be substantial relative to the uncertainty in the mean (the first bracketed term) when researchers have used a small sample size with a highly reliable dependent variable in a repeated-measures study that yielded a substantial SMD. For example, in a crossover with 10 subjects and a moderate SMD (~1.0), the SE^2 due to the first bracketed term could be $\sim 0.20^2 = 0.04$, while that due to the second bracketed term would be $1.0^2/(2 \times 10) = 0.05$.

Another problem with the use of the SMD is the assumption that the dependent variable is normally distributed.[33] It is unclear how the nature and extent of departure from normality will result in difficulties in interpretation of the magnitude thresholds of the SMD.

Finally, a problem with meta-analysis of SMDs is bias arising from the fact the SMD and its SE are not independent. Lin and Oloe[17] showed that the bias is trivial (~ -0.10, in standardized units) for small studies (total sample size of 20) and moderate SMDs (0.80) when Equation (13) is used. Replacing the SMD in this equation with the simple mean of the SMD across all the studies reduces the bias to 0.00 for the special case investigated by Lin and Aloe[17] (the SDs in the two groups differ only because of sampling variation, and their combination is used to standardize). Simulations (not included here) show that this strategy eliminates bias more generally (when pre-intervention or external SDs are used to standardize, when heterogeneity is large, and when a meta-regression is used to account for a group modifying effect with small and large means). The accuracy of the degrees of freedom for the SE of the SMD (Equation 14) with the Lin and Aloe correction could also be investigated, but inaccuracy here will almost certainly make negligible difference to the standardized confidence intervals of individual study estimates in forest plots or tables.

7 | ALTERNATIVES TO SMD FOR META-ANALYSIS

Given all the problems with the SMD, meta-analysts should consider alternatives to combining mean effects from studies with disparate measures. Here we describe three, and the possible use of standardization to evaluate magnitude of the mean effect *after* meta-analysis.

7.1 | Differences between means expressed as factor effects

Mean effects (differences, changes, or differences in the changes) of physiological, biochemical and biomechanical measures that have only positive non-zero values without a theoretical upper bound (eg, concentrations, metabolic rates, measures of performance and size) are likely to be more uniform across different populations and settings when expressed as percents or factors, and such effects are invariant for measures that differ only by a scaling constant. Meta-analysis of the logarithm of such effects expressed as factors will therefore result in less heterogeneity than meta-analysis of raw mean effects, and the resulting meta-analyzed mean and heterogeneity can be back-transformed by simple exponentiation to a mean factor effect and factor SD (and then expressed in percent units as 100*factor-100, if desired). This approach has been promoted previously in the guise of meta-analysis of response ratios[34] or ratios of means[14,35] of ratio-scale measures. An option for deriving meta-analyzed response ratios via log transformation is available in R's metafor.[23(p58)] The Cochrane handbook[14] mentions use of percent effects as an option in section 6.5.1.4 but omits mention of the need to convert percent effects to factors and to log transform. Meta-analyses of response ratios were infrequent (1.7%) in our survey (Table 2), but they may be more popular in ecology, the field in which the seminal work was published by the originator, Hedges.[34]

Ideally, authors of each study will have analyzed their data via log transformation and presented the mean effect and confidence limits as factors (response ratios) or percents. Any percent effects and confidence limits should be converted to factors (given by 1 + x/100, where x is the percent effect or limit) before log-transformation. The SE in log units can then be derived from the log-transformed confidence limits or from the *P* value for the effect. Our simulations of meta-analysis of log-transformed factor or percent effects using the mixed model in SAS (Data S2) include an option for meta-analyzing

effects where all authors have used log transformation: the data in the simulations, which are generated initially as log-transformed values, are analyzed as such and therefore do not result in transformation bias. Back-transformation produces unbiased estimates of the mean factor effect and heterogeneity, both of which can be expressed in percent units, if desired. Use of the log of the factor SE obviates the need for special formulae developed by Lajeunesse[36] for the SE of mean effects representing the ratio of the means of groups that are not independent. If mean effects for two or more groups in some studies are included in the meta-analysis (eg, control and experimental groups; changes from baseline at several time points), between- and within-study heterogeneities need to be estimated with appropriate covariance structures developed by Lajeunesse[36] (see also Program 3 in our simulations).

Unfortunately, analysis of raw data is the norm, and meta-analysis of the resulting log-transformed factor effects produces underestimation of the mean effect and factor heterogeneity when study sample sizes are small.[37] We have previously derived the SE in raw units from the confidence limits or *P* value, then expressed the mean effect and SE as factors (1 + x/M, where x is the mean effect or SE, and M is the reference, baseline, or control mean) before log transformation and meta-analysis (eg, Reference 30). In the course of developing our simulations, we have found that using the t statistic for the raw mean effect to estimate the SE for the log-transformed factor effect (by dividing the log-transformed factor effect by the *t* statistic) results in substantially less bias than use of the log-transformed factor SE (Data S2): the downward bias is practically negligible (factor bias of ~0.99, or ~1.0%) when one or more of the mean effects and factor SDs (between-subject SD, error of measurement, heterogeneity) are factors of up to 1.30 (30%), with sample sizes of 10 in each study and 10 studies in the simulated meta-analysis; with sample sizes of several hundreds, factor effects and SD of up to 2.0 (100%) result in negligible bias.

Meta-analysts opting for meta-analytic models that include effect modifiers or additional random effects should consider using similar simulations to investigate the possible biases in mean and heterogeneity for their particular set of study estimates and sample sizes. Alternatively, a mean of log-transformed factors weighted by sample size rather than the inverse of error variance is unbiased,[38] as confirmed in our simulations, but this method has limitations: it has not been adapted to meta-analyses that include fixed-effect modifiers in meta-regressions; it does not permit inclusion of more than one random effect to represent heterogeneity arising from multiple estimates within studies; and the weighting does not account for differences in error of measurement between studies.

Care is needed with log transformation of factor effects when dependent variables in different studies are related by non-linear transformations. For example, in studies of endurance performance, power (P), speed (S), distance (D), and duration (T) are related by $P = S^n = (D/T)^n$, where *n* varies from 1.0 to 3.0, depending on the mode of exercise. Authors report mean effects usually on only one of these variables. Changes in S, D or T can be expressed as changes in P for meta-analysis, which after log transformation is achieved by multiplying the changes by *n* (for S and D) and by $-n$ (for T).

Variables already representing a percent of a parameter other than the mean, such as body fat as percent of body mass, have finite lower and upper bounds of >0% and <100%, and mean effects on such variables would not usually be more uniform when expressed as factors. The mean percent effects should therefore be meta-analyzed without log or other transformation in the first instance. Log transformation should be considered, if it makes the effect more uniform by removing a positive linear modifying effect of the reference, control or pre-intervention mean value in a meta-regression. When the mean values in different studies cover a wide range, especially if some are close to 0% or 100%, this modifying effect may be non-linear; expressing the means as proportions (p) and applying the logit transformation (log(p/[1-p]) before calculating the mean effect for subsequent meta-analysis may make the effect more uniform and is worth investigating in future simulations of meta-analyses.

7.2 | Mean effects of psychometrics rescaled 0-100

For psychometric dependent variables representing perceptions, attitudes or behaviors assessed with different Likert or visual-analog scales, rescaling to a range of 0-100 rather than log transformation is appropriate for meta-analysis (eg, Reference 39); the common metric is then percent of maximum possible range, and mean effects of ±10, ±30, ±50, ±70, and ± 90% of the range seem reasonable as thresholds for small, moderate, large, very large, and extremely large. These thresholds are probably too large for dimensions of the psyche derived by combining multiple correlated Likert items and rescaling. The approach described next (Section 7.3) would work for such variables, if there are known clinically relevant thresholds for the different measures in the different settings; otherwise, the 0-100 rescaled measures should be meta-analyzed, then the magnitude of the effects should be assessed by standardization, as described in Section 7.4. However, when the mean of a Likert-based measure is close to the minimum or maximum possible value relative to the

SD (the difference between the mean and the minimum or maximum is less than ~3SD), the resulting non-normality of the scores makes interpretation of the SMD problematic. The logit transformation, as described above, needs investigating as a possible solution to this problem.

7.3 | Mean effects rescaled with minimum important differences

For psychometrics representing combinations of different kinds of variable in different settings, an approach that avoids SMDs has been suggested: a rescaling, in which each mean effect and its SE are divided by the practitioner-established minimum clinically important difference in the psychometric in the given setting, then meta-analyzed.[40] A value for the rescaled effect of 1 then represents the threshold for clinical importance; we suggest thresholds of 3, 6, 10, and 20 for moderate, large, very large and extremely large for the rescaled effect, reflecting the spacing of the thresholds for standardized effects.[2] The Cochrane handbook[14] cites this approach in section 6.5.1.4 as an option for combining any mean effects; indeed, it solves all the problems of standardization. The approach was not used in any of the meta-analyses we reviewed.

7.4 | Standardization of mean effects after meta-analysis

In the absence of any scale for clinically important raw, percent, factor, or psychometric mean effects, standardization can be used *after* meta-analysis to assess the magnitude of the meta-analyzed mean, using the methods described in Sections 7.1 and 7.2. Different SDs representing differences between subjects in different settings can be used to standardize; for example, a mean effect might be negligible in settings with a large SD but substantial in those with a small SD. For inferential evaluation of the resulting SMD, its SE $SE_{\Delta Stz}$ is given by a formula similar to Equation (13), again derived by differentiating Equation (1):

$$SE^2_{\Delta Stz} = SE^2_{\Delta}/SD^2_{Stz}/SD_{Stz})^2/(2n_{Stz}), \tag{15}$$

where Δ is the meta-analyzed mean effect, SE_{Δ} is its SE, SD_{Stz} is the standardizing SD, n_{Stz} is its effective sample size, and Δ/SD_{Stz} is the SMD. The degrees of freedom for $SE_{\Delta Stz}$ is given by Equation (14). The meta-analyzed heterogeneity also needs to be standardized and evaluated inferentially by accounting for its uncertainty. If the heterogeneity variance is ν (sometimes presented as τ^2) and its SE is SE_{ν}, then the standardized heterogeneity variance is $\nu/SD_{Stz}{}^2$, and its SE $SE_{\nu Stz}$ is given by differentiation of this expression and first-order approximation as

$$SE^2_{\nu Stz} = SE^2_{\nu}/SD^4_{Stz} + (2/(n_{Stz} - 1))\nu^2/SD^4_{Stz}. \tag{16}$$

The sampling distribution of ν is assumed to be normal.

A standardizing SD taken from a large cohort study or derived by pooling SDs from several studies of similar populations is likely to have a sample size large enough for the contributions of uncertainty in the SD to Equations (15) and (16) to be negligible compared with that of the meta-analyzed mean and heterogeneity, respectively; there would also be no need for adjustment for small-sample bias, and the standardized lower and upper confidence limits of the SMD and heterogeneity (expressed as an SD, τ) would be simply the meta-analyzed limits divided by SD_{Stz}.

If log-transformed factor effects have been meta-analyzed, the underlying assumption is that the original data are log-normally distributed, or at least would be closer to a normal distribution after log transformation. It therefore seems reasonable to convert the standardizing SD to a factor and to log-transform it for standardizing the log-transformed mean effect and heterogeneity. Simulations in SAS (Data S3) show that use of the log-transformed factor SD substantially underestimates the SD of the log-normal data, even for relatively small factor SD, resulting in overestimation of the SMD. The simulations provide a correction factor as a cubic function of the raw SD expressed as a fraction of the raw mean, with coefficients modified by the inverse of the sample size. The factor corrects the SMD for true factor SD of up to 4.0, with an error of up to ~5% for a sample size of 10, and up to ~2.0% for sample sizes of 15 to 1000. The resulting corrected SMD is automatically adjusted for small-sample bias. The correction factor (CF) is given by the following equation:

$$CF = 1/\left(1.027 + (0.346 + 5.62/n)X - (0.173 + 4.57/n)X^2 + (0.021 + 2.76/n)X^3\right), \tag{17}$$

where X is the raw SD divided by the raw mean, and n is the sample size. Data S3 includes a program modified to derive the correction factor as a cubic for a single given sample size; the error in the correction factor is then negligible. The standard error of the adjusted SMD is given by multiplying Equation (16) by CF^2.

A simpler alternative to correcting the SMD derived with the log of a factor SD is to convert the meta-analyzed factor or percent mean effect and heterogeneity to standardized values for assessment of their magnitudes in a given setting, by using the reference mean in that setting to convert the factor or percent values to raw values, then dividing by the raw reference SD. Unfortunately, the resulting SMD is biased, and the bias varies with sample size, SD, and SMD in a manner that did not lend itself to empirical adjustment (Data S3, Program 1). In any case, this method does not solve the problem of interpretation of the magnitude of the SMD when the raw data are non-normal (eg, when the reference SD is similar in magnitude to the reference mean). We do not recommend this method.

If the standardizing SD is for a dimension of the psyche derived by combining Likert-scale items, the SD of the underlying construct is inflated by error of measurement of the mean of the items. This error should be removed by multiplying the observed SD by the square root of Cronbach's alpha, since Cronbach's alpha is defined as the true variance divided by the observed variance (eg, Reference 41). Other things being equal, instruments with different numbers of items should then have similar SDs.

8 | CONCLUSION

In this tutorial we have reasoned that the most appropriate SD for standardizing mean effects is the between-subject SD, free of technical or measurement error, in a reference group, control group, or pre-intervention condition. Use of any other SD, especially the SD of change scores in interventions, can produce substantial bias. Standardization even with an appropriate SD is problematic, and we advise against it. Where possible, meta-analysts should therefore use one of the three alternatives to standardization we have described: log transformation for factor (response ratio) and percent effects, rescaling psychometric data to percent of maximum possible range, or rescaling with minimum clinically important differences. Standardization with one or more appropriate SDs is also an option *after* meta-analysis, but it should be used to assess magnitude in chosen populations only when there is no known clinically important difference.

Our modest sample of recent meta-analyses shows that standardization is a popular approach to combine mean effects from controlled trials and crossovers, but the use of pre-intervention SDs to standardize is uncommon. Readers can trust the magnitudes of mean effects in existing meta-analyses of controlled trials and crossovers, provided that standardization was performed with pre-intervention SDs. The magnitudes will be biased low, when the meta-analysts have failed to remove technical or measurement error from the standardizing SD, but the bias will be negligible for measures with sufficiently high reliability over the time frame of the intervention. If standardization was performed with anything other than pre-intervention SDs, or if it is unclear which SDs were used, readers should not cite the meta-analysis or act on its outcome, unless they can make a reasonable correction to the meta-analyzed magnitude. Many published meta-analyses will need to be discounted and updated.

Developers of meta-analysis software should aim to minimize the harm. The use of inappropriate SD in the meta-analysis of standardized mean effects will persist until the creators of meta-analysis packages provide the appropriate default and other options with documentation that includes clear explanations. In our view, they should allow, and recommend as a last resort, standardizing with the appropriate between-subject SD in the study data or with a suitable between-subject SD sourced elsewhere, and they should promote approaches other than standardization to combine disparate mean effects in a meta-analysis.

Meantime, we urge editors and, in particular, reviewers of manuscripts reporting meta-analyzed SMDs to recommend or even insist on re-analysis with one of the other methods outlined here. To ensure transparency and reproducibility, reviewers and editors should also insist that authors provide adequate detail about their meta-analytic methods; following the advice of the International Committee of Medical Journal Editors: "Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results."[42] Additionally, authors should provide supplemental files of datasets and processing steps to enhance transparency of the approach used in the meta-analysis.

ACKNOWLEDGEMENTS

The authors have no conflicts of interest to declare. The data that supports the findings of this study are available in the supplementary material of this article. Open access publishing facilitated by Massey University, as part of the Wiley - Massey University agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

Data available in article supplementary material.

ORCID

David S. Rowlands https://orcid.org/0000-0003-0017-1515

REFERENCES

1. Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. 1st ed. Hillsdale, NJ: Lawrence Erlbaum; 1969.
2. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc*. 2009;41(1):3-13.
3. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat*. 1981;6(2):107-128.
4. Glass GV, McGaw B, Smith ML. *Meta-Analysis in Social Research*. Beverly Hills, CA: SAGE; 1981.
5. Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res*. 1976;5(10):3-8.
6. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol Methods*. 1996;1(2):170-177.
7. Becker BJ. Synthesizing standardized mean-change measures. *Br J Math Stat Psychol*. 1988;41(2):257-278.
8. Cumming G. *The New Statistics – Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York, NY: Routledge/Taylor & Francis Group; 2012.
9. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol*. 2003;56(5):395-407.
10. Goulet-Pelletier J-C, Cousineau D. A review of effect sizes and their confidence intervals, part I: the Cohen's d family. *Tutor Quant Methods Psychol*. 2018;14(4):242-265.
11. Morris SB. Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *Br J Math Stat Psychol*. 2000;53(1):17-29.
12. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods*. 2002;7(1):105-125.
13. Gibbons RD, Hedeker DR, Davis JM. Estimation of effect size from a series of experiments involving paired comparisons. *J Educ Stat*. 1993;18(3):271-279.
14. Higgins JPT, Li T, Deeks JJ. Chapter 6: choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3. Cochrane; 2022. Accessed March 14, 2023. https://training.cochrane.org/handbook/current/chapter-06#section-6-5
15. Borenstein M. Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. New York, NY: Russell Sage Foundation; 2009:221-235.
16. Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
17. Lin L, Aloe AM. Evaluation of various estimators for standardized mean difference in meta-analysis. *Stat Med*. 2021;40(2):403-426.
18. Luo Y, Funada S, Yoshida K, Noma H, Sahker E, Furukawa TA. Large variation existed in standardized mean difference estimates using different calculation methods in clinical trials. *J Clin Epidemiol*. 2022;149:89-97.
19. Hopkins WG. Individual responses made easy. *J Appl Physiol*. 2015;118:1444-1446.
20. Nunes EA, Colenso-Semple L, McKellar SR, et al. Systematic review and meta-analysis of protein intake to support muscle mass and function in healthy adults. *J Cachexia Sarcopenia Muscle*. 2022;13(2):795-810.
21. Willoughby DS, Stout JR, Wilborn CD. Effects of resistance training and protein plus amino acid supplementation on muscle anabolism, mass, and strength. *Amino Acids*. 2007;32(4):467-477.
22. Oertzen-Hagemann V, Kirmse M, Eggers B, et al. Effects of 12 weeks of hypertrophy resistance exercise training combined with collagen peptide supplementation on the skeletal muscle proteome in recreationally active men. *Nutrients*. 2019;11(5):1072.
23. Viechtbauer W. Metafor: meta-analysis package for R. Version 3.4–0. 2021 https://cran.r-project.org/web/packages/metafor/metafor.pdf
24. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
25. Stata. *Stata Meta-Analysis Reference Manual. Release 17. Statistical Software*. College Station, TX: StataCorp LLC; 2023. Accessed March 14, 2023. https://www.stata.com/manuals/meta/index2.html
26. RevMan. *Review Manager (RevMan)*. [Computer program]. Version 5.4.1. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration. https://training.cochrane.org/online-learning/core-software/revman; 2022.
27. CMA. Comprehensive Meta Analysis (CMA) Version 3.0. Accessed March 14, 2023 https://www.meta-analysis.com/downloads/Meta-Analysis%20Manual%20V3.pdf
28. Yang M. *A Review of Random Effects Modelling in SAS (Release 8.2)*. London: Centre for Multilevel Modelling, Institute of Education, University of London; 2003.
29. Rowlands DS, Kopetschny BH, Badenhorst CE. The hydrating effects of hypertonic, isotonic and hypotonic sports drinks and waters on central hydration during continuous exercise: a systematic meta-analysis and perspective. *Sports Med*. 2022;52(2):349-375.
30. Patten RK, Boyle RA, Moholdt T, et al. Exercise interventions in polycystic ovary syndrome: a systematic review and meta-analysis. Systematic review. *Front Physiol*. 2020;11:606.
31. Hopkins WG. Sample-size estimation for various inferential methods. *Sportscience*. 2020;24:17-27.

32. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics*. 1946;2(6):110-114.
33. Bonett DG. Confidence intervals for standardized linear contrasts of means. *Psychol Methods*. 2008;13(2):99-109.
34. Hedges LV, Gurevitch J, Curtis PS. The meta-analysis of response ratios in experimental ecology. *Ecology*. 1999;80(4):1150-1156.
35. Friedrich JO, Adhikari NKJ, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. *BMC Med Res Methodol*. 2008;8(1):32.
36. Lajeunesse MJ. On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology*. 2011;92(11):2049-2055.
37. Lajeunesse MJ. Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*. 2015;96(8):2056-2063.
38. Bakbergenuly I, Hoaglin DC, Kulinskaya E. Estimation in meta-analyses of response ratios. *BMC Med Res Methodol*. 2020;20(1):263.
39. Machado LA, Kamper SJ, Herbert RD, Maher CG, McAuley JH. Analgesic effects of treatments for non-specific low back pain: a meta-analysis of placebo-controlled randomized trials. *Rheumatology (Oxford)*. 2009;48(5):520-527.
40. Johnston BC, Thorlund K, Schünemann HJ, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. *Health Qual Life Outcomes*. 2010;8:116.
41. Hammond S. Using psychometric tests. In: Breakwell GM, Hammond S, Fife-Schaw C, Smith JA, eds. *Research Methods in Psychology*. 3rd ed. London: SAGE Publications Ltd; 2006:182-209.
42. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. 2011. Accessed April 13, 2024 https://www.icmje.org/about-icmje/faqs/icmje-recommendations/

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hopkins WG, Rowlands DS. Standardization and other approaches to meta-analyze differences in means. *Statistics in Medicine*. 2024;1-17. doi: 10.1002/sim.10114