
Article:
Gallardo Gómez, Daniel, Pedder, Hugo, Welton, Nicky J. et al. (2 more authors) (2024) Variability in meta-analysis estimates of continuous outcomes using different standardization and scale-specific re-expression methods. Journal of Clinical Epidemiology. 111213. ISSN 0895-4356

https://doi.org/10.1016/j.jclinepi.2023.11.003

Variability in meta-analysis estimates of continuous outcomes using different standardization and scale-specific re-expression methods

Daniel Gallardo-Gómez[1,2], Hugo Pedder[3], Nicky J. Welton[3], Kerry Dwan[4], Sofia Dias[5]

[1]Department of Physical Education and Sports, Faculty of Education, University of Seville, 41013 Seville, Spain.

[2]Epidemiology of Physical Activity and Fitness Across the Lifespan Research Group (EPAFit).

[3]Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, BS82PS, UK.

[4]Liverpool School of Tropical Medicine, Pembroke Place Liverpool, L3 5QA, UK.

[5]Centre for Reviews and Dissemination, University of York, YO10 5DD, UK.

Corresponding author: Daniel Gallardo Gómez; +34 633 236 562; dggomez@us.es; Department of Physical Education and Sport, Faculty of Educational Sciences, University of Seville, 41013, Seville, Spain.

Abstract

Objective To explore the impact of using different data standardization and scale-specific re-expression methods (i.e., processes to convert standardized data into scale-specific units) in meta-analyses using standardized mean differences (SMDs).

Study design and setting We used data assessed by the Short Physical Performance Battery and the Barthel Index from a meta-analysis of randomized controlled trials which synthesized evidence of physical activity effectiveness on the functional capacity of hospitalized older adults. We standardized the data using study-specific pooled SDs, an internal, and an external SD references. Bayesian meta-analyses were performed for each method to compare the posterior distributions of the meta-analysis parameters. Posterior estimates were re-expressed into scale-specific units applying different methods established in the Cochrane guidelines.

Results Meta-analysis estimates depend on the used standardization method. Analyses including data standardized using the largest SD reference presented lower estimates with less uncertainty in both scales. The method applied for re-expressing SMDs into scale-specific units impacted in their posterior clinical interpretation. The most similar results across models were obtained when using the same SD reference to standardize and re-express data.

Conclusion Different data standardization methods yielded different meta-analysis estimates on the SMD scale. To avoid the introduction of bias, the use of a single scale-specific SD reference to standardize data is recommended, and instead of study-specific pooled sample SDs. Meta-analysis software packages may therefore change their default methods to allow

this method by a single scale-specific SD. To re-express the SMDs into scale-specific units, we suggest the application of the same SD reference that was used for data standardization.

Keywords: Meta-analysis; standardized mean difference; effect size; evidence synthesis; standardization; clinical interpretation

Running title: Variability in meta-analysis estimates of continuous outcomes

Word count: 3,060 words

What is new?

- Different standardization methods showed different meta-analysis estimates.
- A single scale standard deviation reference is recommended for data standardization.
- The medical interpretation of an effect measure depends on how it is re-expressed.
- We suggest the use of the same reference for standardization and re-expression.

1 | Introduction

Standardized mean differences (SMD) are effect size measures commonly used in meta-analysis of continuous outcomes, which allow pooling of data on the same outcome reported on different measurement scales [1]. Usually, this effect metric is expressed as Cohen's *d* [2] or Hedges' *g* [3] in social and medical sciences. The most common method used for standardization –and the default in most of meta-analysis packages in R [4] (e.g., *metafor* [5]; or *esc* [6])– consists of dividing the mean difference (MD) between treatment and control groups by the pooled sample standard deviation (SD) in each study at a post-treatment time point [7]. However, this process is not recommended by evidence synthesis methodological guidelines because of the inclusion of potential sources of heterogeneity by using numbers highly dependent on a plethora of uncontrolled factors (e.g., bias or individual prognostic factors), which differ even across included studies using the same scale. Thus, the National Institute for Health and Care Excellence (NICE) Guidelines Technical Support Unit (GTSU) Guidelines Methodology Document 2 (GMD2) recommends the use of a fixed scale-specific SD reference, which could be 1) an existing SD from an external reference population [8], or if that is not possible, 2) an internal reference standard such as the average of pooled SDs at baseline for each scale [9]. However, no study has compared these three methods in the same dataset to contrast and explain any resulting differences.

Despite the popularity of SMDs in meta-analyses, an additional major concern flagged by methodologists is the clinical interpretability of these estimands. Some general rules of thumb exist (e.g., SMD ≥ 0.2 and < 0.5 is a small effect; SMD ≥ 0.5 and < 0.8 is a moderate effect; and SMD ≥ 0.8 is a large effect; [2]), but most methodologists believe that such interpretations are problematic because 'patient' importance of a finding is context-dependent and not amenable to generic statements [7]. To address this issue, a possible solution is to re-express the pooled SMDs to scale-specific MD units [7]. In practice, the most common conversion is carried out by multiplying the SMD estimates by the same SD reference used for standardization. Nonetheless, Cochrane methodological guidelines recommend the use of a specific type of SD reference depending on whether or not the scale of interest is included in the meta-analysis. Yet, epidemiologists highlighted that this back-conversion process may imply the possibility of "severe distortions" that can even reverse the magnitude of the effect estimates [1]. This study therefore aimed to shed light, through an illustrative example, on (a) the impact of using different standardization methods to obtain SMDs in meta-analysis model parameters' estimates, and (b) the potential clinical implications of using different methods for re-expressing the standardized estimates.

2 | Illustrative example: effectiveness of physical activity on the functional capacity of hospitalized older adults

We used a dataset from a systematic review with meta-analysis that included randomized controlled trials which applied a physical activity-based intervention to improve the functional capacity of hospitalized older adults aged 50 or over [10]. Several scales were included in this study, but for illustrative purposes, we focus here only on data collected on (1) the Short Physical Performance Battery (SPPB), and (2) the Barthel Index (BI) scales.

The SPPB is an objective assessment tool for evaluating lower extremity functioning in older adults. The possible scores range from zero (worst performance) to 12 (best performance) points. The BI measures functional disability in 10 activities of daily living (ADL) by quantifying patient performance. The possible scores for this scale range from zero ("total" dependency) to 99 ("slight" dependency).

3 | Methods

3.1 | Data

Our datasets included contrast-based data as mean differences between physical intervention and control groups and their associated standard errors. We computed the standardized mean differences of these values by:

– Study-specific 1: Dividing the MDs and SEs by the pooled sample SD of each study at the pre-intervention time point.

– Study-specific 2: Dividing the MDs and SEs by the pooled sample SD of each study at the post-intervention time point.

– Internal reference: Using an internal SD reference standard calculated as the average of the pooled SDs at baseline for each scale, which corresponds to 2.42 for the SPPB and 16.62 for the BI scales.

– External reference: Using an existing SD from an external reference population obtained by scientific literature search of studies that were considered to represent the patient population of the trials included in the meta-analysis [9]. A retrospective cohort study [11] including 375 older adults admitted due to an acute illness was considered to represent the patient population of the trials. For SPPB, the external SD reference at baseline corresponds to 3.14; and for BI corresponds to 25.39.

Scales and SD references details are displayed in Table 1. In addition, SPPB and BI complete datasets with all transformations are shown in Table 2. All the standardization procedures are fully detailed in the Supplementary Material 1.

Table 1. Scale details and SD values for standardization

Scale	Number of studies	Score range	Study-specific pooled sample SDs*	Study-specific pooled sample SDs**	Internal reference	External reference
SPPB	7	0 to 12	1.62 to 2.71	2.22 to 3.60	2.42	3.14
BI	5	0 to 99	9.86 to 26.00	9.86 to 25.00	16.27	25.39

Notes. SPPB: Short Physical Performance Battery; BI: Barthel Index. *Displayed values are the ranges of the calculated study-specific pooled sample SDs for each study/scale at pre-intervention. **Displayed values are the ranges of the calculated study-specific pooled sample SDs for each study/scale at post-intervention

Table 2. Data for physical activity interventions effect on the functional capacity of hospitalized older adults

Outcome	Study	Scale-specific units		Study-specific pooled sample SDs at pre-time point		Study-specific pooled sample SDs at post- time point		Internal SD reference		External SD reference*	
		MD	SE	SMD	SE	SMD	SE	SMD	SE	SMD	SE
Short Physical Performance Battery	Campo, 2019	2.000	0.291	0.889	0.129	0.900	0.131	0.826	0.120	0.637	0.093
	Casas-Herrero, 2022	0.500	0.487	0.194	0.189	0.194	0.189	0.206	0.201	0.152	0.148
	Kitzman, 2021	–0.100	0.290	–0.037	0.107	–0.037	0.107	–0.041	0.120	–0.030	0.088
	Martinez-Velilla, 2019	1.900	0.374	0.730	0.144	0.528	0.104	0.784	0.155	0.576	0.113
	Martinez-Velilla, 2021	1.900	0.493	0.761	0.197	0.761	0.197	0.784	0.203	0.576	0.149
	Martinez-Velilla, 2022	0.100	0.417	0.062	0.258	0.034	0.142	0.041	0.172	0.030	0.126
	Ortiz-Alonso, 2020	–0.600	0.356	–0.222	0.132	–0.213	0.127	–0.248	0.147	–0.182	0.108
Barthel Index	Casas-Herrero, 2022	0.490	1.861	0.050	0.189	0.050	0.189	0.030	0.114	0.111	0.423
	de Morton, 2007	4.000	3.812	0.154	0.147	0.163	0.155	0.246	0.234	0.909	0.866
	Martinez-Velilla, 2019	1.000	1.768	0.061	0.107	0.059	0.104	0.061	0.109	0.227	0.402
	Martinez-Velilla, 2021	12.500	3.354	0.735	0.197	0.735	0.197	0.768	0.206	2.841	0.762
	Martinez-Velilla, 2022	3.910	1.839	0.050	0.189	0.301	0.141	0.030	0.114	0.111	0.423

Note. Columns with suffixes *_ext*, *_int*, or *_default* are SMDs and their SEs. *****External SD references were extracted from Urquiza et al. [10]; SPPB SD reference = 3.14, and BI SD reference = 25.39.

3.2 | Meta-analysis models

Bayesian multilevel random-effects meta-analysis models were performed to estimate the expectation of the posterior predictive distribution of our estimands (i.e., pooled MDs or SMDs, and between-study heterogeneity). Once meta-analysis models were fitted, we re-expressed the standardized estimates to scale-specific units applying different methods to compare how subsequent treatment recommendations could vary between them. We multiplied the standardized estimates by:

– Method 1: The same SD reference used for data standardization. To preserve the number of posterior draws between models, we used the most applicable method to re-express study-specific 1 and 2 models' estimates using one SD reference value corresponding to a weighted internal SD reference (explained in the Method 2).

– Method 2: A weighted SD reference calculated as the average of pre-intervention SD values across all intervention groups of all studies that used the selected instrument (Supplementary Material 2). Cochrane methodological guidelines state that this is a reasonable option when the scale of interest is included in the meta-analysis [7].

– Method 3: An external SD reference from a representative observational study. Cochrane methodological guidelines state that this option should be used when the scale of interest is not included in the meta-analysis [7].

Our models were fitted using a vague normal prior distribution for the pooled relative effect parameter with mean zero and scale (SD) 100 ($N(0, 100)$). To incorporate information about the between-study heterogeneity (SD) parameter, we defined minimally informative prior distributions, truncating to only positive values ($\tau \sim Half\ Cauchy(10)$), whose 95% prior density lies between 0 and 127.06 for both outcomes, which is very wide to the range of observed effects.

3.2.1 | Model implementation

Models were run using 4 chains with 5000 iterations per Markov Chain Monte-Carlo (MCMC) chain; 1000 iterations were discarded to ensure iterations were only saved once chains had converged. The thinning rate, which ensures optimal monitorization, saved results for 1 in every 2 iterations per chain. To assess the convergence and overall validity of our models, we first checked the values of the Potential Scale Reduction Factor (PRSF) of the estimated parameters, which should be smaller than 1.01 [12]. Second, we conducted a posterior predictive check by comparing simulated random draws from our model and the observed data, and if the model has converged and fitted the data well, the density of the replications should be roughly similar to the one of the observed data (Supplementary Material 3).

All the statistical analyses were performed in R software [4]. We used the *brms* package [13, version 2.18.0] to perform Bayesian meta-analysis models; the *tidybayes* package [14, version 3.0.2] to integrate Bayesian modelling into tidy data; and the *ggridges* [15, version 0.5.4] and *ggplot2* [16, version 3.3.6] packages for data plotting and visualization. The code and data required to reproduce the results presented in this manuscript are available through public repository access (link: https://github.com/dgalgom/Variability-in-meta-analysis-estimates-of-continuous-outcomes).

3.3 | Model fit

We used the Deviance Information Criterion (DIC) to compare meta-analysis models' fit that used the same likelihood. Lower DIC values indicate lower deviance, and thus, better model fit. Although there is no established consensus, differences of more than 10 might rule out the

model with the higher DIC, and between 5 and 10 are substantial [17]. Comparison between models' fit is shown in the Supplementary Table 1.

3.4 | Sensitivity analysis

Considering the popularity of frequentist statistics in evidence synthesis research, we also conducted all the meta-analysis models under a frequentist approach to possibly ease the interpretation of the results. The meta-analyses conducted under this approach assumed random-effects models. Taken into account the data (i.e., continuous outcome data) and to ease the replicability of these results, we used as between-study heterogeneity estimators the restricted maximum likelihood and DerSimonian-Laird estimators.

4 | Results

4.1 | Standardized meta-analysis estimates

4.1.1 | Short Physical Performance Battery

Models that meta-analyzed mean differences standardized by study-specific pooled sample SD references (at pre- and post-intervention time points), and by an internal SD reference showed similar SMD estimates (study-specific 1 = 0.35, 95% CrI –0.22 to 0.87; study-specific 2 = 0.30, 95% CrI –0.21 to 0.80; internal reference = 0.33, 95% CrI –0.22 to 0.86). The external reference model, which standardized data by using the highest SD value and yielded the best model fit (Supplementary Table 1), presented a smaller relative effect with less associated uncertainty (SMD = 0.24, 95% CrI –0.15 to 0.63). In a similar way, the external reference model showed a smaller between-study SD estimate (τ = 0.28, 95% CrI 0.02 to 0.85) than the rest of the models (study-specific 1 = 0.37, 95% CrI 0.02 to 1.09; study-specific 2 = 0.36, 95% CrI 0.02 to 1.04; internal reference = 0.38, 95% CrI 0.02 to 1.11). All standardized mean estimates and 95% CrI are plotted in Figure 1.

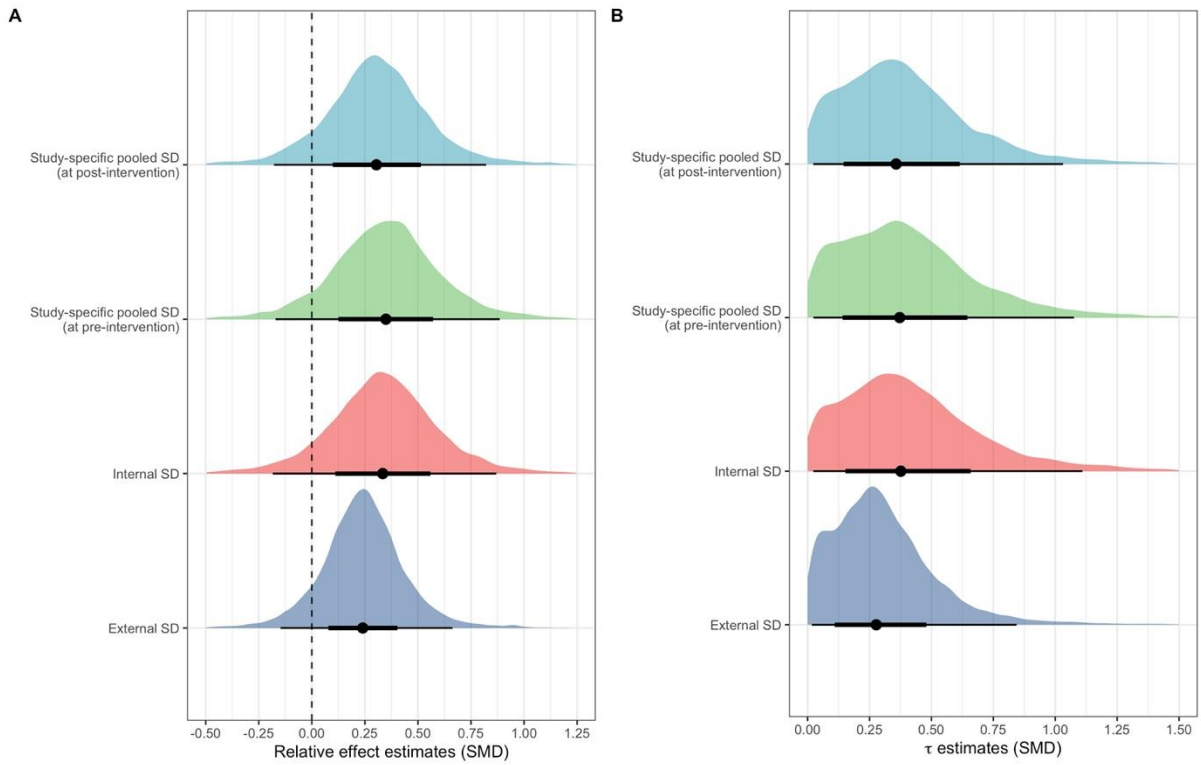

Figure 1. A: Standardized mean estimates and 95% CrI of SPPB outcomes. B: Heterogeneity estimates and 95% CrI of SPPB outcomes.

4.1.2 | Barthel Index

Models that included data standardized by study-specific pooled sample SD references (pre- and post-intervention time points), and by an internal SD reference showed similar relative effects for physical activity (study-specific 1 = 0.25, 95% CrI –0.28 to 0.75; study-specific 2 = 0.25, 95% CrI –0.30 to 0.76; internal reference = 0.24, 95% CrI –0.28 to 0.75). The external reference model, which had the highest SD and presented the lower DIC value (Supplementary Table 1), presented a smaller SMD estimate with less uncertainty (SMD = 0.14, 95% CrI –0.22 to 0.49). The external reference model showed a smaller between-study SD estimate (τ = 0.14, 95% CrI 0.01 to 0.76) than the rest of the models (study-specific 1 = 0.25, 95% CrI 0.01 to 1.07; study-specific 2 = 0.25, 95% CrI 0.01 to 1.11; internal reference = 0.26, 95% CrI 0.01 to 1.09). All standardized mean estimates and 95% CrI are plotted in Figure 2.

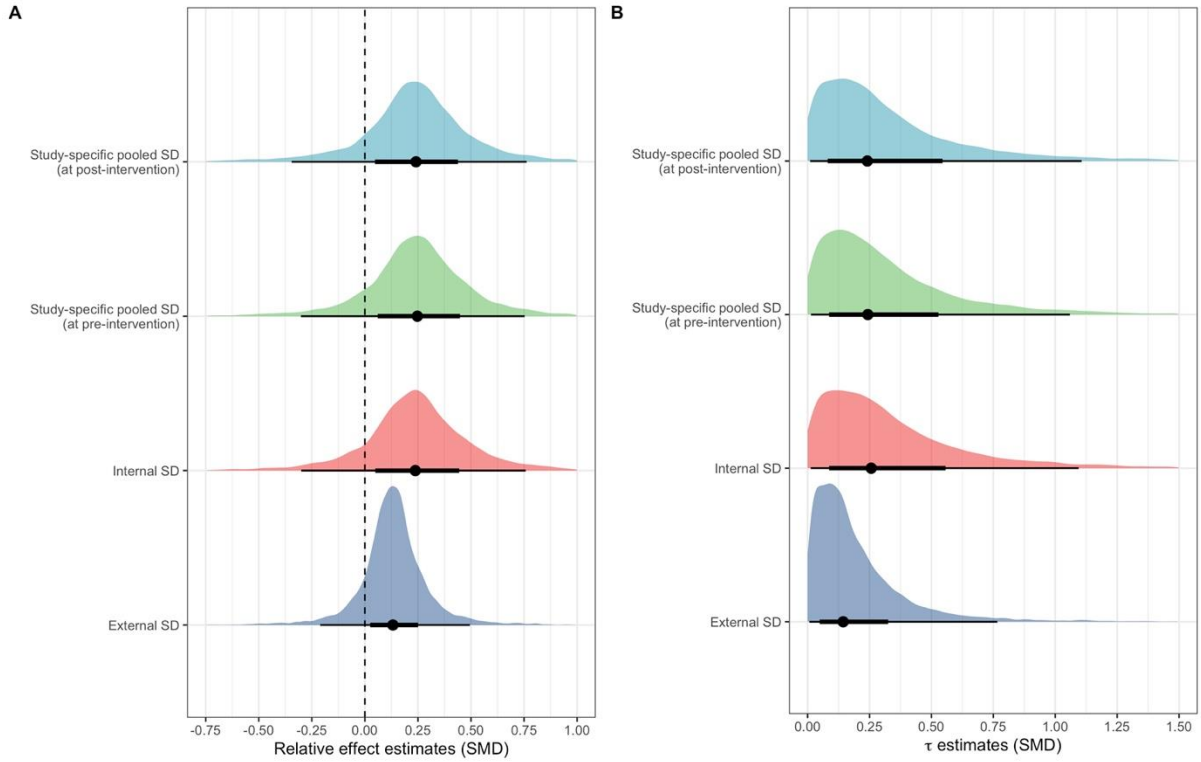

Figure 2. A: Standardized mean estimates and 95% CrI of BI outcomes. B: Heterogeneity estimates and 95% CrI of BI outcomes.

4.2 | Re-expressing SMD estimates into MD units

4.2.1 | Short Physical Performance Battery

The application of method 1 to re-express standardized estimates yielded similar relative effects with great interval overlap between the models (range = 0.75 to 0.85), including the scale-specific MD estimate (MD = 0.82, 95% CrI –0.50 to 2.16). Methods 2 and 3 presented wider ranges of potential effects (method 2 range = 0.59 to 0.87; method 3 range = 0.75 to

1.10). The range of between-study SD estimates using the method 1 was smaller (range = 0.87 to 0.92) than those obtained through methods 2 (range = 0.69 to 0.94) and 3 (range = 0.88 to 1.19), including the heterogeneity estimate from the MD model (τ = 0.92, 95% CrI 0.06 to 2.79). Original (i.e., unstandardized) and converted MD estimates, and 95% CrI are depicted in Figure 3. Numerical data can be found in the Supplementary Table 2.

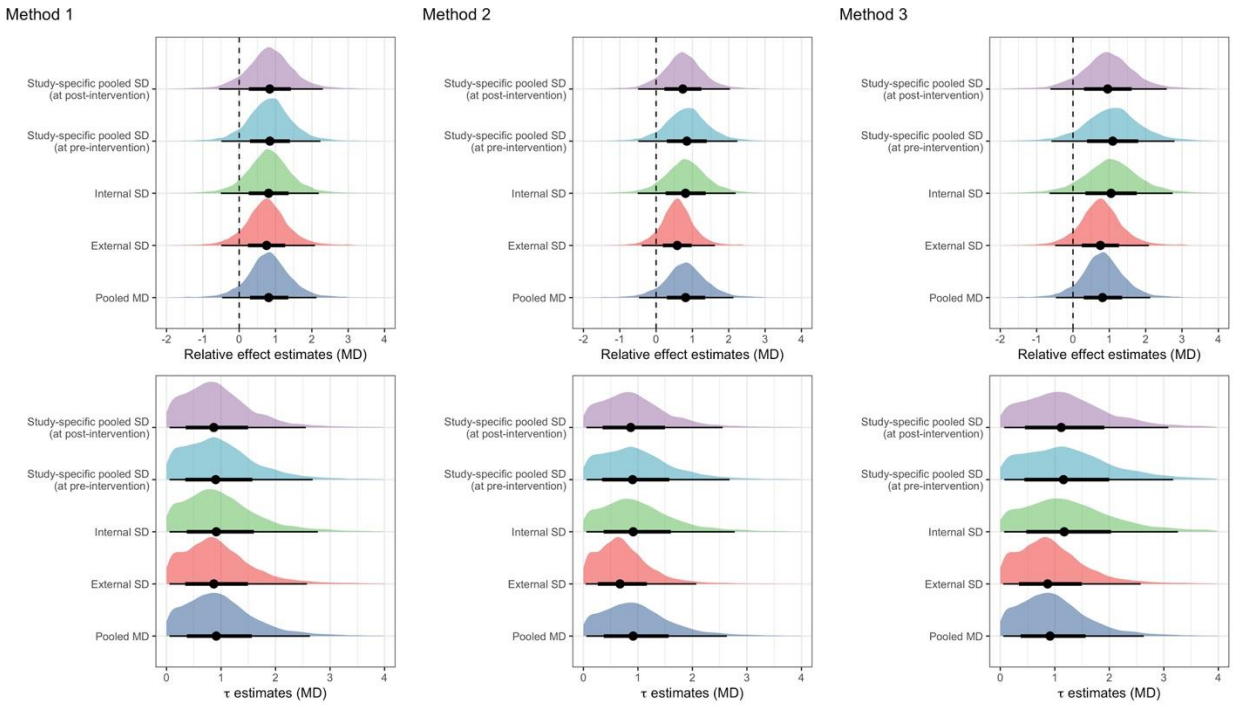

Figure 3. Converted mean difference and heterogeneity estimates using different re-expressing methods for SPPB outcomes. Method 1: using the same SD reference for standardization. Method 2: using a weighted SD reference calculated as the average of pre-intervention SD values. Method 3: using an external SD reference from a representative observational study. Pooled MD refers to original MD values (i.e., no standardization).

4.2.2 | Barthel Index

Re-expressing standardized estimates applying method 1 showed great overlap between converted scale-specific estimates (range = 3.42 to 4.04), including the scale-specific MD estimate (MD = 3.75, 95% CrI –2.15 to 10.20). Methods 2 and 3 presented greater differences between mean effect estimates (method 2 range = 2.32 to 4.27; method 3 range = 3.42 to 6.30). The between-study SD estimate from the MD model (τ = 3.37, 95% CrI 0.17 to 12.10) was lower than the transformed standardized heterogeneity estimates using the method 1 (range = 3.70 to 4.24). Methods 2 and 3 presented more disperse between-study SD estimates ranges (method 2 range = 2.51 to 4.49; method 3 range = 3.70 to 6.62). Original and transformed MD estimates, and 95% CrI are plotted in Figure 4. Numerical data can be found in the Supplementary Table 2.

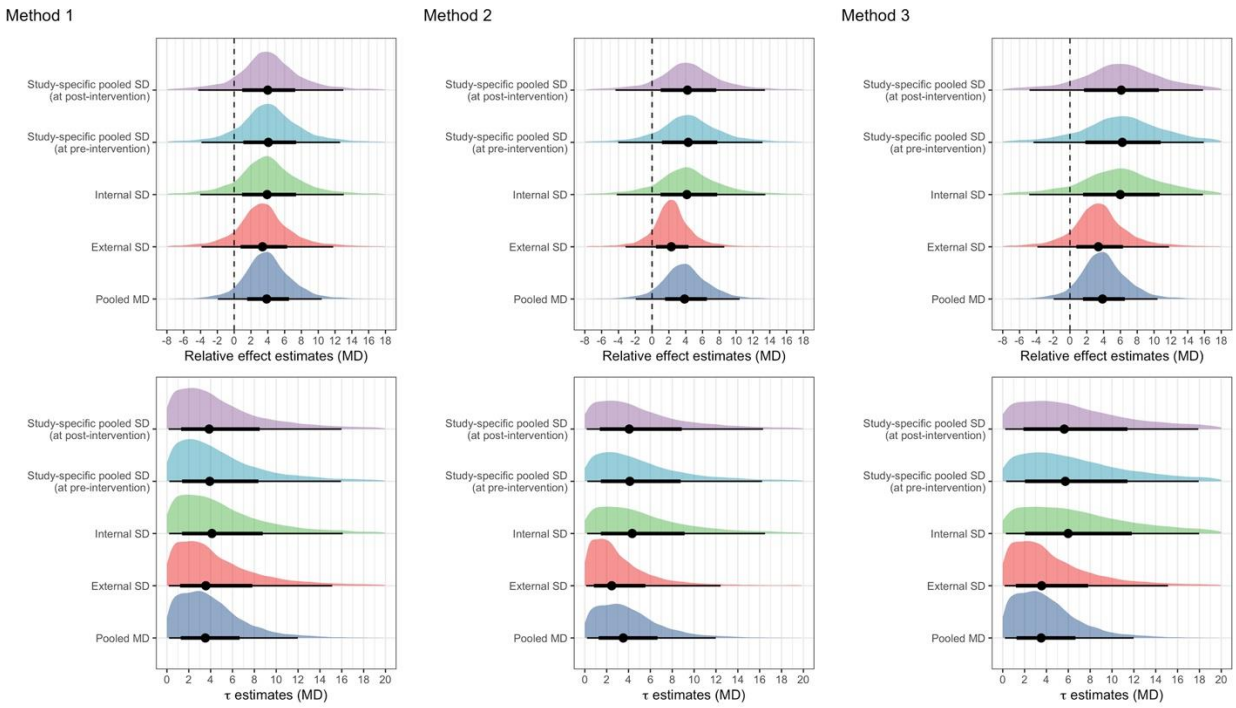

Figure 4. Converted mean difference and heterogeneity estimates using different re-expressing methods for BI outcomes. Method 1: using the same SD reference for standardization. Method 2: using a weighted SD reference calculated as the average of pre-intervention SD values. Method 3: using an external SD reference from a representative observational study. Pooled MD refers to original MD values (i.e., no standardization).

4.3 | Sensitivity analyses

Meta-analysis estimates under a frequentist approach showed very similar results compared with our main models. Results for SPPB data is depicted in Table 3, and for BI data in Table 4. Plots with all meta-analysis estimates are presented in the Supplementary Material 4.

Table 3. Short Physical Performance Battery meta-analysis results under a frequentist approach

Model*	**Estimator**	**Estimate (95% CI)****	**τ (95% CI)**	**Re-expressed estimate (95% CI)**	**Re-expressed τ (95% CI)**
Study-specific 1	DL	0.34 (–0.08 to 0.76)	0.46 (0.25 to 1.03)	0.95 (–0.28 to 2.19)	1.27 (0.74 to 2.94)
Study-specific 2	DL	0.30 (–0.09 to 0.70)	0.41 (0.24 to 0.94)	1.07 (–0.24 to 2.38)	1.44 (0.79 to 3.22)
Internal	DL	0.33 (–0.09 to 0.75)	0.45 (0.25 to 1.02)	1.05 (–0.27 to 2.37)	1.42 (0.80 to 3.19)
External	DL	0.24 (–0.06 to 0.55)	0.33 (0.18 to 0.74)	0.76 (–0.20 to 1.71)	1.03 (0.58 to 2.31)
Pooled MD	DL	0.81 (–0.21 to 1.83)	1.04 (0.60 to 2.38)		
Study-specific 1	REML	0.30 (–0.09 to 0.70)	0.40 (0.24 to 0.92)	0.95 (–0.28 to 2.19)	1.26 (0.74 to 2.90)
Study-specific 2	REML	0.34 (–0.08 to 0.76)	0.43 (0.25 to 0.97)	1.07 (–0.24 to 2.38)	1.34 (0.77 to 3.04)
Internal	REML	0.33 (–0.09 to 0.75)	0.43 (0.25 to 0.98)	1.05 (–0.28 to 2.37)	1.35 (0.78 to 3.09)
External	REML	0.24 (–0.06 to 0.55)	0.31 (0.18 to 0.71)	0.76 (–0.20 to 1.71)	0.97 (0.57 to 2.24)
Pooled MD	REML	0.81 (–0.21 to 1.83)	1.04 (0.60 to 2.38)		

Note. DL: DerSimonian-Laird; REML: Restricted Maximum Likelihood. *Pooled MD refers to original MD values (i.e., no standardization). **Estimates from study-specific, internal, and external models are presented as SMDs, and the estimate from Pooled MD model is presented as MD.

Table 4. Barthel Index meta-analysis results under a frequentist approach

Model*	Estimator	Estimate (95% CI)**	τ (95% CI)	Re-expressed estimate (95% CI)	Re-expressed τ (95% CI)
Study-specific 1	DL	0.24 (–0.09 to 0.57)	0.19 (0.01 to 0.74)	6.06 (–2.27 to 14.39)	4.82 (0.35 to 18.85)
Study-specific 2	DL	0.24 (–0.09 to 0.57)	0.19 (0.02 to 0.75)	6.14 (–2.29 to 14.57)	4.90 (0.47 to 19.06)
Internal	DL	0.23 (–0.11 to 0.56)	0.19 (0.05 to 0.77)	5.72 (–2.28 to 14.32)	4.77 (1.23 to 19.46)
External	DL	0.12 (–0.06 to 0.31)	0.10 (0.03 to 0.42)	3.16 (–1.59 to 7.90)	2.63 (0.66 to 10.74)
Pooled MD	DL	3.67 (–1.84 to 9.18)	3.06 (0.78 to 11.38)		
Study-specific 1	REML	0.24 (–0.09 to 0.57)	0.20 (0.00 to 0.79)	6.08 (–2.27 to 14.44)	5.00 (0.00 to 20.15)
Study-specific 2	REML	0.24 (–0.09 to 0.58)	0.20 (0.00 to 0.80)	6.17 (–2.29 to 14.63)	5.10 (0.00 to 20.28)
Internal	REML	0.24 (–0.11 to 0.58)	0.21 (0.02 to 0.83)	5.88 (–2.85 to 14.60)	5.43 (0.62 to 21.09)
External	REML	0.13 (–0.06 to 0.32)	0.12 (0.01 to 0.46)	3.24 (–1.57 to 8.06)	3.00 (0.32 to 11.63)
Pooled MD	REML	3.77 (–1.83 to 9.36)	3.48 (0.40 to 11.01)		

Note. DL: DerSimonian-Laird; REML: Restricted Maximum Likelihood. *Pooled MD refers to original MD values (i.e., no standardization). **Estimates from study-specific, internal, and external models are presented as SMDs, and the estimate from Pooled MD model is presented as MD.

5 | Discussion

This study has shown through an illustrative example that meta-analysis models including the same data but using different standardization methods yielded different results. We also found that the application of different re-expression methods to convert standardized meta-analysis results in scale-specific units resulted in distinct estimates. Altogether, the findings revealed in this study highlight the importance of explaining very carefully the methodological process undertaken in a meta-analysis that requires data standardization (i.e., the studies included in the meta-analysis used different scales to measure the same construct), and how treatment effects and between-study heterogeneity estimates are re-expressed to clinically interpret them.

The variations presented using different standardization methods are rooted in the mathematical fact of dividing the effect measure (e.g., mean difference) and their standard error by different SD references, whose values may be scale-dependent: especially, scales with a large score range are more likely to have a higher SD reference value. It could be reflected into lower relative effects measured as SMDs, and lower between-study SD values, both with a tighter uncertainty distribution. In our example, the Barthel Index (score range from 0 to 99) showed lower SMDs with less associated uncertainty than the Short Physical Performance Battery (score range from 0 to 12). Because of the strong association between SMD magnitude and the SD reference used for standardization, traditionally accepted cut-off points to interpret the impact of a SMD as small, moderate, or large have been demonstrated potentially controversial. Importantly, that could lead to incorrect interpretations when a meta-analysis of SMDs would favor the treatment group against the control (i.e., 95% CrI would not include the zero), since it would be difficult to know whether the detected treatment effect is due to the uncertainty shrinkage caused by a large SD reference for standardization (particularly problematic when scales have a wide score range), which could be also detected at study-specific level (Supplementary Material 5), or real effectiveness of the analyzed intervention.

Most meta-analysis statistical software packages contribute to the extended belief that the unique existing and available standardization method is based on the use of study-specific pooled sample SDs, which is the most common method by default. Capitalizing on the current guidelines and the results presented in this study, we recommend adding alternative options to allow a single-specific SD (e.g., user-specified or internally calculated) to be used instead and consider changing packages defaults to an internally calculated scale-specific SD, with options for user-specified values to be given. Specifically, the GTSU GMD2 suggests the preferred use of an external SD reference for standardization as the most feasible option to mitigate the introduction of potential biases. Hence, a repository including different SD references collected from large representative studies for different populations from more general groups (e.g., older adults) to more specific ones (e.g., older adults admitted due to an acute health condition) would be useful for meta-analyses 'standardization' (i.e., homogeneity), and thus, evidence synthesis replicability.

Due to the potential issues flagged by methodologists regarding the medical interpretation of SMDs, Cochrane methodological guidelines recommend using natural units of measurement to present the results of meta-analyses of continuous outcomes using these effect measures. In our example, we observed that if the same SD reference was used for data standardization and scale-specific re-expression (i.e., method 1), very similar relative effects and between-study heterogeneity were obtained in both scales. Nevertheless, if we re-expressed SMDs using a familiar instrument as Cochrane recommendations state, a great variability existed

depending on which method was used to standardize the data. Therefore, we suggest the use of the same SD reference for standardization and re-expressing procedures.

Overall, analyses on the original MD scales are preferable but standardization may be required for meaningful synthesis, when different scales are used to measure the same outcome in different studies. However, nowadays, a huge number of meta-analyses are using SMDs as effect measure, sometimes unnecessarily, and making treatment recommendations based on their estimated 'effect sizes'. In this study we explored the implications of using different standardization methods, highlighting the latent biases that a SMD measure could suffer. In addition, we also note the potential implications for conclusions on the clinical relevance when SMDs are re-expressed into a meaningful MD scale. Yet, one limitation of this study is that, in a real meta-analysis, we would pool all available evidence from different scales to estimate a pooled effect, which would be transformed into scale-specific effects by multiplying it by the corresponding scale-specific SD reference (an organization chart of this procedure is depicted in the Supplementary Material 6). The distortions in pooled effects from these combinations are expected to follow a similar pattern as those shown here, with maybe some potential for additional increases in estimated heterogeneity. Additionally, the study protocol was not officially registered. Finally, we observed these phenomena in one of the simplest subsets of evidence synthesis techniques, pairwise meta-analysis, in which a unique treatment group is compared against a control group. Future research could appear in more complex evidence synthesis methods like network meta-analyses, and even if any of their assumptions (e.g., consistency) may be compromised depending on the approaches used for data standardization and effect measure re-expression.

6 | Conclusions

Meta-analyses of mean differences from different scales are feasible only after applying standardization procedures that achieve a standardized measurement unit such as the widely used standardized mean differences. Through an illustrative example, we demonstrated that the use of different standardization methods resulted in different pooled effect size and between-study heterogeneity estimates. In line with methodological guidelines, we suggest the use of a single scale-specific SD for data standardization (i.e., external reference or an internally calculated one) instead of study-specific pooled sample SDs. The defaults for calculating SMDs in most meta-analysis software packages may therefore be changed to allow a single scale-specific SD to be used. As transformed scale-specific unit should be used to evaluate the clinical effectiveness of a treatment, we recommend the use of the same SD reference for data standardization and scale-specific re-expression. Altogether, this work sheds light on current issues with meta-analyses of continuous outcomes that used standardized mean differences as effect measure, and makes recommendations to address them.

7 | Competing interests

Authors have no competing interests to declare.

8 | Authors' contribution

DGG and SD conceptualized the study design. DGG and HP conducted the formal statistical analyses with critical input from SD; DGG, HP, and SD designed and conducted data plotting

and visualization. DGG, HP, and SD drafted the manuscript with critical input from NJW and KD. All authors approved the final version of the manuscript to be submitted.

9 | Funding

This work was supported by a predoctoral teaching and research fellowship via I+D+i Research Program of the University of Seville, Spain (PIF20/VI PPIT-US), and an associated grant for short stays abroad for the development of the University of Seville's own I+D+i Research Programme (VIIPPIT-2023-EBRV).

10 | References

[1] Ades AE, Lu G, Dias S, Mayo-Wilson E, Kounali D. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. Res Synth Methods 2015;6:96–107.
[2] Cohen J. Statistical Power Analysis for the Behavioral Sciences. Academic Press; 2013.
[3] Hedges LV. Distribution Theory for Glass's Estimator of Effect size and Related Estimators. J Educ Behav Stat 1981;6:107–28.
[4] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2022.
[5] Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. J Stat Softw 2010;36:1–48.
[6] Lüdecke D, Lüdecke MD, Calculator'from David BW. Package "esc" 2017.
[7] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons; 2019.
[8] Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials 2011.
[9] Daly C, Anwer S, Welton NJ, Dias S, Ades AE. NICE Guidelines Technical Support Unit n.d.
[10] Gallardo-Gómez D, del Pozo-Cruz J, Pedder H, Alfonso-Rosa RM, Álvarez-Barbosa F, Noetel M., et al. Optimal dose and type of physical activity to improve functional capacity and minimize adverse events in acutely hospitalised older adults: A systematic review with dose-response network meta-analysis of randomised controlled trials. BJSM 2023.
[11] Urquiza M, Fernández N, Arrinda I, Sierra I, Irazusta J, Rodríguez-Larrad, A. Nutritional status is associated with function, physical performance, and falls in older adults admitted to geriatric rehabilitation: a retrospective cohort study. Nutrients 2020;12:2855–69.
[12] Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci 1992;7:457–72.
[13] Bürkner P-C. Advanced Bayesian Multilevel Modeling with the R Package brms. arXiv [statCO] 2017.
[14] Kay M. tidybayes: Tidy data and geoms for Bayesian models. R package version 3.0. 0 2021.
[15] Wilke CO. ggridges: ridgeline plots in "ggplot2." R Package Version 05 n.d.
[16] Wickham H. ggplot2. Wiley Interdiscip Rev Comput Stat 2011;3:180–5.
[17] Spiegelhalter DJ, Best NG, Carlin BP, Van de Linde A. Bayesian Measures of Model Complexity and Fit (with Discussion). Journal of the Royal Statistical Society (Series B) 2002;64:583–616