Reference Standardizers for Effect Size Calculation in Meta-Analysis: An Empirical Investigation in Psychological Treatment Research & Development of a Searchable Database

Bernat Bruno Fabian, Clara Miguel, Katrin Jansen, Paula Kuper, Toshi A. Furukawa, Pim Cuijpers, Annemieke van Straten, Mathias Harrer

Status: 2026-02-23

Summary

This project investigates how using reference standard deviations instead of study-specific SDs affect effect size estimates in meta-analyses of psychological interventions. In most meta-analyses, effect sizes (like Hedges' *g*) are calculated by dividing the mean difference between groups by the pooled SD from each trial. However, this method can be biased due to inconsistent or missing SDs, especially in small-sample RCTs. Our goal is to assess if replacing these with *reference SDs*, pooled from similar studies using the same outcome scales, improves reliability and reduces heterogeneity in meta-analytic results. We will compile a catalogue of reference SDs from the Metapsy database. These reference SDs will be calculated overall and within subgroups (e.g., by age group, administration type, comparator type, diagnosis method). We will compare meta-analytic results using: (1) original study-specific SDs, (2) reference SDs, and (3) subgroup-specific reference SDs.

Description

In meta-analysis, the most used measure of effect size is the standardized mean difference (SMD), which enables the combination of results across studies that assess the same outcome using different measurement (Gallardo-Gómez et al., 2024; Nakagawa et al., 2015). SMDs are typically expressed as Cohen's *d*, Hedges' *g*, or Glass's *δ*. In the context of social and medical sciences, the prevailing method of calculating these metrics involves dividing the mean difference (MD) between the treatment and control groups by the pooled standard deviation (SD) within each study (Gallardo-Gómez et al., 2024), typically at post-test. This approach assumes that the SDs are sufficiently similar across studies using the same scale, thereby allowing for meaningful comparison and aggregation of effect sizes (Downing et al., 2025; Gallardo-Gómez et al., 2024).

However, variability in study-specific SDs and missing data can complicate this standardization process, necessitating alternative approaches to mitigate its impact. SDs often vary across studies even if they use the same instrument measuring the outcome (Friedrich et al., 2008). These differences could be due to sample characteristics, population heterogeneity, or sampling error, introducing statistical noise and potentially biasing effect size estimates (Downing et al., 2025). Moreover, many studies simply do not report SDs which makes the calculation of SMDs even more difficult. Lastly, in studies with small sample sizes, as is often the case in randomized controlled trials, standard deviation estimates are likely to be highly imprecise and potentially biased (Wan et al., 2014). Therefore, addressing the challenges posed by inconsistent or missing SDs is essential to ensure accurate and unbiased estimation of SMDs. To overcome these limitations, researchers have explored alternative standardization strategies that do not rely solely on study-specific SDs.

To address the previously mentioned challenges, alternative standardization methods have been proposed. One such approach involves dividing the mean difference by a scale-specific constant either derived from an external reference drawn from a normative population or using an internal reference by pooling the SDs across all studies employing the same scale (Downing et al., 2025;

Gallardo-Gómez et al., 2024). This strategy using large databases could help reduce artefacts caused by between-study variability in SDs and enhances the comparability and interpretability of the resulting standardized effect sizes (Downing et al., 2025).

Building on these considerations, the present study aims to advance the standardization of effect sizes in psychological treatment research by developing a comprehensive catalogue of reference SDs and accompanying software tools. All reference standardizers will be calculated from living databases curated by the Metapsy collaboration (metapsy.org; Harrer et al., 2025), covering 12 different mental health problems, and ensuring that versioned updates can be released as new evidence accumulates. These resources are intended to support the consistent application of reference SD standardization methods across diverse psychological outcome measures in psychological treatment research. In an empirical investigation, we also seek to evaluate the impact of using reference SDs on pooled effect size estimates and between-study heterogeneity across more than twelve mental health conditions. Ultimately, we also aim to explore which study-level variables are associated with higher or lower trial-level standard deviations post-intervention.

Research Questions and Hypotheses

1. How does the use of reference SDs, as opposed to study-specific SDs, affect pooled effect size estimates in meta-analyses of psychological treatments?

Hypothesis:
Using reference SDs will result in more consistent indicated τ^2 by and stable pooled effect size estimates measured by the standard error.

2. How does reference SD standardization affect estimates of between-study heterogeneity?

Hypothesis:
Reference SD standardization will reduce between-study heterogeneity.

Secondary Research Questions

We also plan to explore if several study-level characteristics predict systematic differences in SD across outcomes: (i) age group (adolescents, young adults, adults, elderly), (ii) administration type (self-report vs. clinician-administered), (iii) diagnosis method (clinical diagnosis vs. cut-off score), (iv) type of comparator (treatment-as-usual, wait-list, psychoeducation, active comparator, other), (v) country or income region, (vi) risk of bias (low vs. not low), (vii) type of analysis (intention to treat vs. completers analysis) and (viii) baseline log SD-ratio. See “Statistical Analyses”.

Study type

Meta-analysis

Blinding

No blinding is involved in this study.

Study Procedures

In the first phase of the project, we will generate a catalogue of reference standard deviations (SDs) for commonly used outcome instruments by pooling log-transformed post-intervention SDs across studies using the same scale. Where sufficient data are available (≥3 studies), we will also derive fine-grained reference SDs for subgroups defined by characteristics such as age group, administration type, comparator type, diagnosis method, and risk of bias. In the second phase, we will use these reference SDs to calculate SMDs and compare meta-analytic results across three conditions: (i) using study-specific SDs, (ii) using default reference SDs, and (iii) using fine-grained subgroup-specific reference SDs. We will assess how the choice of standardizer affects pooled effect size estimates and between-study heterogeneity. Finally, we will conduct multiple meta-regression analyses to assess variation in SD sizes.

Randomization
Because we will not conduct a randomized trial but a meta-analysis, we will not randomize.

Data collection procedures and Explanation of existing data
Analyses will draw on data collected by Metapsy, a comprehensive meta-analytic database of RCTs examining the effects of psychological interventions on various mental health problems (Cuijpers et al., 2019; Harrer et al., 2025). We will include Metapsy databases focusing on twelve mental health problems: unipolar depression, panic disorder, SAD, GAD, specific phobias, PTSD,OCD, BPD, prolonged grief disorder, problem gambling, psychotic disorders (including schizophrenia/psychosis, schizophreniform dis-order, schizoaffective disorder, delusional disorder), and suicidality. Methodological details for all of these databases are provided in the respective documentation entries (Cuijpers, Miguel, Harrer, Plessen, Ciharova, Ebert, & Karyotaki, 2023; Mc Glanaghy et al., 2023; Papola, 2023a, 2023b; Pfund et al., 2023b; Setkowski et al., 2023b; van Ballegooijen et al., 2023). For the present study, we will use the most recent available version of each database. All datasets are publicly available via the project website and GitHub repository, which also includes additional metadata: github.com/metapsy-project. A comprehensive description of the database methodology is presented by Cuijpers, Karyotaki, Ciharova, et al. (2019) and Harrer et al. (2025). Screening and data extraction were performed independently by two researchers, with any disagreements resolved through discussion with a third reviewer.

Full search strategy details can be found at docs.metapsy.org/databases. All databases included in Metapsy are documented on a dedicated database webpage, organized by indication. Each database has a corresponding GitHub repository that stores the current and previous versions, along with a metadata folder containing release-specific information. Databases are also indexed via Zenodo and assigned a DOI that always resolves to the latest version; earlier versions have unique DOIs for citation. Each database is assigned a shorthand identifier, which can be used in tools like metapsyData for streamlined access in R (docs.metapsy.org/databases).

Inclusion and exclusion criteria
We will use the same eligibility criteria as used by Harrer et al. (2025), but with the additional inclusion of unguided treatment options. Therefore, to be eligible for the database, a study has to be (a) an RCT that compares (b) any psychological intervention (guided/unguided), (c) for any adult age or target group with (d) any comparison group. Only trials published in peer-reviewed journals are considered. “Gray literature” (i.e., dissertations or other publications without formal peer-review) was considered in previous versions, but it is not systematically included in the latest versions of the database (Harrer et al., 2025).

Sample size
It is not clear how many trials will be included in the meta-analysis, but we expect up to 1,000 RCTs or more in total that meet our inclusion criteria.

Sample size rationale
Since this is a meta-analysis, we can only rely on existing and available data.

Stopping rule
A stopping rule cannot be used in this meta-analytic project.

Manipulated variables
Variables that are used in this meta-analytic study include the variables that are routinely collected as part of our main meta-analytic project (characteristics of the participants, of the interventions, of the study).

Main variables

- Pooled standard deviation (SD) of each study (at post and follow-ups, if available). Outcome instruments can be any validated clinician- or patient-reported (PROM) rating scale resulting in a continuous score, so that SDs can be meaningfully calculated in both groups.
- Small-sample bias corrected standardized mean differences (Hedges' g), between intervention group and comparator, computed using (i) trial-specific pooled SDs, and (ii) reference standard deviations for the chosen instrument.
- For the exploratory analysis, we will employ log-variation ratios as the primary outcome. For each study, we will calculate the log-variation ratio between the observed ("empirical") SD and the estimated reference SD, derived from all studies using the same instrument (see "Statistical Analysis").

Independent/Additional variables

In addition to scale-specific reference SDs, we will also pool SDs across subgroups of studies using the same measurement scale, with stratification by age group (adolescents, young adults, adults, elderly), administration type (self-report vs. clinician-administered), risk of bias (high vs. low), diagnosis method or severity (clinical diagnosis vs. cut-off score), type of comparator (e.g., treatment as usual, waitlist, psychoeducation, active comparator, or other), presence of comorbidity, and intention to treat vs. completers analysis. Each subgroup must include a minimum of three studies. The goal of this process is to derive fine-grained reference SDs for inclusion in the catalogue (see "Dissemination & Online Tools" for further details)

Predictors

Our meta-regression model will examine the following study-level predictors of variability in standard deviations: participant age group (adolescents, young adults, adults, elderly), administration type (self-report vs. clinician-administered), diagnosis method (clinical diagnosis vs. cut-off score), type of comparator (treatment-as-usual, wait-list, psychoeducation, active comparator, or other), recruitment method, country of origin, income region (classified as high, upper middle, lower middle, or low), cultural region (Western vs. non-Western), risk of bias (low vs. not low), and baseline log SD-ratio (calculated using the post-intervention reference standardizer). The goal of employing meta-regression analysis is to examine whether there are systematic factors influencing higher or lower outcome variability across instruments and disorders.

These meta-regression analyses cannot account for potentially systematic differences between studies using different instruments. For example, certain instruments may be used more frequently or exclusively in trials targeting specific populations, such as older adults, or specific indications. If such factors influence post-randomization pooled standard deviations, the resulting reference standardizers may in some cases be less representative and may reduce the comparability of effect sizes calculated using different instrument-specific standardizers. We will therefore assess the impact of such selection biases by visually inspecting and reporting the distributions of the study level predictors listed above for each disorder and instrument. Markedly divergent distributions of clinical or baseline characteristics for one instrument compared with others within a disorder category will be examined as indicators of limited representativeness and/or reduced comparability with other reference standardizers.

Query string

Query strings for each living database can be found in their respective documentation entries at docs.metapsy.org/databases.

Data management and sharing

Metapsy databases are officially released from the main branch of their GitHub repositories. Each release follows a versioning system similar to the Semantic Versioning standard. All releases are automatically indexed on Zenodo, which assigns both a database DOI (pointing to the latest version) and a version DOI (specific to each release). Once indexed, the database and metadata are integrated across the Metapsy infrastructure, including documentation, the metapsyData R package, and the

Metapsy API. Citations are automatically updated to reflect the correct year and version. For further details, see docs.metapsy.org/release/.

Software

All analyses will be conducted in R version 4.2.0 or higher. Analyses will mainly be conducted using the metapsyTools package, which was developed specifically for Metapsy-type databases as employed in this study. This package imports functionality from the metafor, meta, and clubSandwich package, among others.

Statistical Analyses

As a first step, eligible outcome data included in the Metapsy databases will be used to create a catalogue of meta-analytic reference standardizers. To this end, we will use inverse variance random-effects pooling models synthesizing bias-corrected versions of the log-transformed pooled SD (s) for each comparison. For each measurement instrument, the following formula is used for the outcome measure *s* and its variance (Nakagawa et al., 2015):

$$\hat{\theta}_k = \log_e(s) + \frac{1}{2(n_1 + n_2 - 2)}$$
$$\hat{V}[\hat{\theta}_k] = \frac{1}{2(n_1 + n_2 - 2)}$$

All log-transformed SDs will be back-transformed to their original scale after pooling. In our main model, all available pooled SDs for post-intervention comparison within a specific study will first be aggregated (i.e., if a study includes more than one post-test assessment, pooled SDs of these multiple assessments will be aggregated). These study-aggregated (independent) pooled SDs will then be pooled across studies. To aggregate outcome measures within studies (e.g., when there are multi-arm trials or multiple valid post-randomization assessments using the same instrument), we will assume an intra-study correlation coefficient of ρ = 0.5.

To assess the robustness of our findings, we will conduct a series of sensitivity analyses. First, we will estimate the pooled SD using a three-level correlated and hierarchical effects (CHE) model, as described by Pustejovsky and Tipton (2022). This model accounts for the dependency of outcome measures within studies and will assume an intra-study correlation of ρ = 0.5. Cluster-robust variance estimation (CRVE) using the CR2 estimator will be applied to guard against model misspecification. Second, we will conduct a meta-analysis excluding statistical outliers, as identified using the "non-overlapping confidence intervals" method, whereby a study is considered an outlier if its 95% confidence interval (CI) does not overlap with the 95% CI of the pooled effect size (Harrer et al., 2021). Third, we will exclude influential cases from the analysis based on the influence diagnostics proposed by Viechtbauer and Cheung (2010). This approach helps identify studies that disproportionately affect the overall results. Fourth, we will conduct two additional meta-analyses: one including only the smallest pooled SD reported within each study, and another including only the largest. Fourth, we will pool effects using a fixed/common-effect model. Finally, if possible, we will estimate the reference SD based solely on studies rated as having a low risk of bias.

These pooling steps will be applied in each indication-specific database when at least 3 studies measuring an instrument of interest are available. If possible, we will also calculate more fine-grained, subgroup-specific reference SD estimates for an instrument. Thus, if at least 3 studies are available, we will also use the same methods as described above to obtain SD estimates conditional on (i) age group (adolescents, young adults, adults, elderly), (ii) administration type (self-report vs. clinician-administered), (iii) diagnosis method (clinical diagnosis vs. cut-off score) and (iv) type of comparator (treatment-as-usual (TAU), wait-list, psychoeducation, active comparator, other etc.).

After compiling the reference SD catalogue, as a second step, we will investigate the impact of using reference versus trial-estimated SD standardizers in meta-analyses of psychological intervention trials. Thus, for each individual database, we will calculate the pooled and subgroup-specific effect of

treatment compared to controls, following the analytic approach described in Harrer et al. (2025). This analysis will be repeated three times: (i) using the original trial-specific standardizers, (ii) using the "default" reference SD obtained for each instrument, and (iii) using the more "fine-grained" subgroup-specific reference SD, if available. To showcase the impact of employing reference SD in meta-analytic psychological treatment research under realistic conditions, we will (i) retain trial-specific SD comparators if the instrument used in the study was available in less than 3 trials (i.e., when no reference SD is available in the catalogue); and (ii) add and report studies to the meta-analysis when their effect sizes can only be obtained using reference standardizers. As a sensitivity analysis, we will also calculate effects when (ii) is omitted. To gauge the impact of using reference SDs, we will then compare the pooled effect size and between-study heterogeneity variance across analyses (2).

Lastly, we will use reference standardizers in the catalogue to determine potential study-level predictors of higher or lower pooled SDs. These analyses will be restricted to studies involving instruments for which reference SDs could be obtained, and will employ the estimated reference SD $\log_e(s_{\text{ref}})$ using all studies employing the same instrument for an indication (i.e., subgroup-specific SD estimates, if available, will not be used). Using $\log_e(s_{\text{ref}})$, the outcome used in these analyses will be the log-variation ratio between reference and "empirical" SD in the study.

$$\hat{\theta}_k = \log_e(s_k) - \log_e(s_{\text{ref}}) + \frac{1}{2(n_1 + n_2 - 2)}$$

The sampling variance of $\hat{\theta}_k$ is given by:

$$\hat{V}[\hat{\theta}_k] = \text{Var}[\log_e(s_k)] + \text{Var}[\log_e(s_{\text{ref}})] - 2\,\text{Cov}[\log_e(s_k), \log_e(s_{\text{ref}})],$$

where the variance of the reference standardizer $\text{Var}[\log_e(s_{\text{ref}})]$ is obtained directly from the meta-analytic model used to estimate the reference SD for the corresponding instrument, and with:

$$\text{Var}[\log_e(s_k)] = \frac{1}{2(n_1 + n_2 - 2)}.$$

Because the s_{ref} is computed as a weighted mean of study-specific log-SDs, a non-zero covariance arises between it and $\log_e(s_k)$ when study k contributes to the reference estimate. Assuming independence of log-SDs across studies and inverse-variance weighting, this covariance simplifies to the product of the normalized (random-effects) meta-analytic weight $w_k^* = (v_k + \tau^2)^{-1} / \sum_{j=1}^{K}(v_j + \tau^2)^{-1}$ of study k and the sampling variance of $\log_e(s_k)$, that is:

$$\text{Cov}[\log_e(s_k), \log_e(s_{\text{ref}})] = w_k^* \frac{1}{2(n_1 + n_2 - 2)}$$

This outcome will then be used in a multiple meta-regression model across indications. The following predictors will be examined: (i) age group (adolescents, young adults, adults, elderly), (ii) administration type (self-report vs. clinician-administered), (iii) diagnosis method (clinical diagnosis vs. cut-off score), (iv) type of comparator (Treatment-as-usual (TAU), wait-list, psychoeducation, active comparator, or other), (v) country/income region, (vi) risk of bias (low, not low), (vii) baseline log SD-ratio (with the post-intervention reference standardizer in the catalogue used as s_{ref}).

Dissemination & Online Tools

To facilitate the application of reference SDs in meta-analyses and trial evaluations of psychological interventions, we plan to make the catalogue developed in this study accessible to the public. First, the final catalogue will be versioned according to the Metapsy data standard (docs.metapsy.org/release/#versioning) and receive a designated documentation entry at docs.metapsy.org/databases. This documentation entry will display the metadata of the catalogue, including the number of included

studies and their references, search details, and other methodological details, along with a download link of the entire catalogue. Furthermore, we aim to enhance the documentation entry by adding more fine-grained search features that allow users to obtain suitable standardizers for their outcome using a graphical user interface online. If feasible, we will also add functionality to the metapsyTools R package (tools.metapsy.org) that allows us to apply reference standardization to our catalogue by default when calculating effect sizes. Via integration into the Metapsy data infrastructure (see Harrer et al., 2025, supplement), we ensure that new versions of the database can easily be re-released as new studies become available, while ensuring backward compatibility using version-specific DOIs.

Transformations
No additional transformations are planned unless specified in the statistical models section above.

Inference criteria
We will use the conventional significance threshold (p=0.05) to determine if differences between groups of studies are statistically significant.

Data exclusion
Studies employing uncommon instruments with fewer than three included studies ($k < 3$) and studies reporting only binary outcome data will be excluded from the analyses.

Funding
No external funding.

Overlapping authorship
We acknowledge that some coauthors may be involved in trials included in the meta-analysis. However, risk of bias assessments are not conducted by study (co-)authors by default, ensuring that these evaluations remain independent. Furthermore, the data analyst for this project has not been involved in any of the analyzed trials.

Conflict of interest
None.

References

Cuijpers, P., Karyotaki, E., Reijnders, M., & Ebert, D. D. (2019). Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression. *Epidemiology and Psychiatric Sciences*, *28*(1), 21–30. https://doi.org/10.1017/S2045796018000057

Cuijpers, P., Karyotaki, E., Ciharova, M., Quero, S., Pineda, B., Muñoz, R., Ph.D., … Rosenström, T. (2022, February 18). A meta-analytic database of randomised trials on psychotherapies for depression. https://doi.org/10.17605/OSF.IO/825C6

Downing, B. C., Welton, N. J., Pedder, H., Mavranezouli, I., Megnin-Viggars, O., & Ades, A. E. (2025). Synthesis of depression outcomes reported on different scales: A comparison of methods for modelling mean differences. *Research Synthesis Methods*. https://doi.org/10.1017/rsm.2025.7

Friedrich, J. O., Adhikari, N. K. J., & Beyene, J. (2008). The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study. *BMC Medical Research Methodology*, *8*. https://doi.org/10.1186/1471-2288-8-32

Gallardo-Gómez, D., Pedder, H., Welton, N. J., Dwan, K., & Dias, S. (2024). Variability in meta-analysis estimates of continuous outcomes using different standardization and scale-specific re-expression methods. *Journal of Clinical Epidemiology*, *165*. https://doi.org/10.1016/j.jclinepi.2023.11.003

Harrer, M., Miguel, C., Van Ballegooijen, W., Ciharova, M., Plessen, C. Y., Kuper, P., Sprenger, A. A., Buntrock, C., Papola, D., Cristea, I. A., De Ponti, N., Baši´cbaši´c, Đ., Pauley, D., Driessen, E., Quero, S., Grimaldos, J., Fernández Buendía, S., Botella, C., Hamblen, J. L., … Cuijpers, P. (2025). *Effectiveness of Psychotherapy: Synthesis of a "'Meta-Analytic Research Domain'" Across World Regions and 12 Mental Health Problems*. https://doi.org/10.5281/zenodo.11116214

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, *6*(2), 143–152. https://doi.org/10.1111/2041-210X.12309

Wan, X., Wang, W., Liu, J., & Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. In *BMC Medical Research Methodology* (Vol. 14). http://www.biomedcentral.com/1471-2288/14/135