Journal of Consulting and Clinical Psychology

ISSN: 0022-006X

2026, Vol. 94, No. 6, 393–399
https://doi.org/10.1037/ccp0001016

COMMENTARY

(Computation of) the Standardized Mean Difference in Clinical Trials Should Be Standardized

Alexander O. Crenshaw
Department of Psychological Science, Kennesaw State University

The standardized mean difference (SMD) is widely used in psychology clinical trials to convert intervention effect estimates from raw units to a common scale, enabling comparison across settings. However, multiple operationalizations of the SMD are possible for clinical trial data, with large potential impacts on estimate magnitudes. In this commentary, I review practices for operationalizing the SMD in the *Journal of Consulting and Clinical Psychology* and other journals that publish results of clinical trials. I find that only one third of studies report sufficient information to determine the operationalization. Among those, the operationalization varies widely, threatening the SMD's utility in providing a common scale. Given the common goals and design elements across many trials, it is feasible to establish a default operationalization for the SMD's standardizer. I outline eight desirable SMD properties and evaluate each option against these properties. The pooled baseline standard deviation meets seven of eight, far more than any alternative, with its main limitation being susceptibility to inflation from restrictive inclusion criteria. I conclude that raw effect estimates should be divided by the pooled baseline standard deviation as a default, one-size-fits-most choice. Adopting this standard will help the SMD meet its promise as a standardized metric, improving comparability of results across different measures, outcomes, studies, and populations. I also provide reporting recommendations to improve transparency and enable recomputation using alternative standardizers when needed.

What is the public health significance of this article?
This commentary reviews practices for standardizing effect estimates in psychology clinical trials. It concludes that practices vary widely, leading to noisy estimates and limiting the benefits of standardization. It concludes with a recommendation to adopt one approach for standardizing treatment effect estimates for most studies.

Keywords: clinical trial, effect size, standardized mean difference, Cohen's *d*, Hedges's *g*

Clinical trials are widely used in psychology to test intervention effects on psychological, behavioral, social, and/or biological outcomes, and these studies are frequently published in the *Journal of Consulting and Clinical Psychology*. A central goal of many clinical trials is to communicate information about the direction and magnitude of treatment effects on key outcomes. It is thus important that effect estimates have commonly understood meanings among researchers, clinicians, and other stakeholders. However, many outcomes commonly tested in psychology trials, such as symptom measures, lack such commonly understood meanings as their scales are often arbitrary. Users must therefore be familiar with the distribution of a particular symptom scale, of which there are hundreds

Pim Cuijpers served as action editor.

Alexander O. Crenshaw https://orcid.org/0000-0002-7453-2398

The present study was not preregistered. Data and the coding manual are available at https://doi.org/10.17605/OSF.IO/CV83U.

The author declares that there are no conflicts of interest. This work was funded by a seed grant and the Dissemination Acceleration Program from the Radow College of Humanities and Social Sciences at Kennesaw State University awarded to Alexander O. Crenshaw. The grant sponsor played no role in study design; the collection, analysis, and interpretation of data; the writing of this article; or the decision to submit this article for publication.

The author thanks Ameesha MacDonald, Kianan Carr, Yeonkuk Woo, Madison Blackmon, Alena Shull, Sophia Pope, and other members of the Couples and Methodology Lab at Kennesaw State University for their time, effort, and conscientiousness in creating the review database and coding articles. The author thanks Danica Kulibert for providing helpful feedback on earlier versions of the article.

Alexander O. Crenshaw played a lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Alexander O. Crenshaw, Department of Psychological Science, Kennesaw State University, 402 Bartow Avenue North West, Kennesaw, GA 30144, United States. Email: acrens13@kennesaw.edu

if not thousands in use, to interpret the magnitude of a given estimate in the scale's raw units.

Estimates of the standardized mean difference (SMD)[1]—such as Cohen's *d*, Hedges's *g*, and Glass's Δ—are widely used to overcome these limitations of arbitrary scales. The SMD is computed by dividing a mean difference (MD) in raw units of the outcome measure (the numerator) by the measure's standard deviation (*SD*; the denominator). The MD component is derived from the trial estimand and may be, for example, the (baseline-adjusted) between-group difference at a posttreatment time point, the between-group difference in change from baseline, or, for an uncontrolled trial, the within-group change from baseline. The issue of the optimal MD estimate has been addressed thoroughly in prior work (e.g., Daly et al., 2021; Feingold, 2013; Higgins et al., 2024; see the Open Science Framework page for a longer list) and is not a focus of this commentary.

Conceptually, the "standardized" part of the term, represented by the denominator, is best understood as simply a unit conversion: Like converting different units of length into centimeters, the SMD converts raw units of a particular measure to *SD* units. This unit conversion enables the interpretation of the magnitude of treatment effects using a single common scale, which in turn enables comparison of intervention effects across different measures, outcomes, studies, and populations. Meta-analyses combining different measures of a construct also require comparability of SMDs to aggregate results (Higgins et al., 2024).

However, clinical trials typically have data structures with multiple possible options for the standardizer (Cumming, 2013; Feingold, 2013; Luo et al., 2022), such as the baseline *SD*, posttreatment *SD*, *SD* pooled across time points, *SD* of change scores, or a model-derived *SD*. Many estimates can also be either computed separately within each treatment arm or subgroup or pooled across treatment arms or subgroups. This proliferation of options creates a problem for the SMD as a standardized metric. If users differ in the standardizer used, the SMD is no longer standardized in any meaningful sense of the word. Moreover, the comparison of SMD results across studies is no longer an apples-to-apples one, leading to noisier estimates of treatment effects and impairing meta-analysis (Hopkins & Rowlands, 2024).

One prior study assessed variability in SMD operationalization in randomized controlled trials (RCTs) in medical fields. Luo et al. (2022) examined 171 SMD operationalizations in medical RCTs and found that 30 used the baseline *SD* for the denominator, 27 used the posttreatment *SD*, nine used the change *SD*, 25 used a model-derived *SD*, 14 used another method, and the method was unclear for 66. This study demonstrated wide variation of SMD operationalization in medical trials, but whether the same is true of psychology trials is unclear.

More comprehensively demonstrated is that the choice of standardizer can greatly impact the magnitude of SMD estimates. Harrer et al. (2025) found that meta-analytic SMD estimates in depression trials varied from 0.65 to 1.24 in one subset of studies and from 0.05 and 0.14 in another subset, depending on the specific SMD operationalization. Similarly, Gallardo-Gómez et al. (2024) found meta-analytic SMD estimates for one outcome varied depending on whether the baseline or posttreatment *SD* was used, though not for another. Ostinelli et al. (2024) noted that baseline *SD*s were smaller than posttreatment *SD*s in a study of different SMD operationalizations in a meta-analysis for depression, which impacts SMD estimates. Furukawa and Leucht (2013) found baseline and posttreatment *SD*s diverged when studies implemented stricter severity thresholds for inclusion. Finally, Luo et al. (2022) conducted perhaps the most thorough analysis by computing the SMD in three or more distinct ways for each individual trial that reported sufficient information to do so, finding the median difference in magnitude between the largest and smallest SMD was 0.30 (range = 0.02–1.61). This value is comparable to proposed values for the smallest worthwhile differences for psychotherapy (e.g., Sahker et al., 2025, Supplement 5). Therefore, individual researchers' choice of standardizer alone, holding all other factors constant, has the potential to influence clinical decision making.

Variability of SMD Operationalizations in Clinical Psychology Trials

Though wide variation in specific SMD operationalization in medical fields has been demonstrated, it is unknown whether this pattern also applies to clinical trials in psychology. To address this gap, my research lab conducted a methodological scoping review of clinical trials published in clinical psychology journals in the years 2020–2023. Based on journals ranked highly in "clinical psychology" on the SCImago Journal & Country Rank database (https://www.scimagojr.com) and that also routinely publish results of clinical trials, we ultimately selected four journals: *Journal of Consulting and Clinical Psychology*, *Depression and Anxiety*, *Journal of Clinical Child and Adolescent Psychology*, and *Psychotherapy and Psychosomatics*. This review was not preregistered.

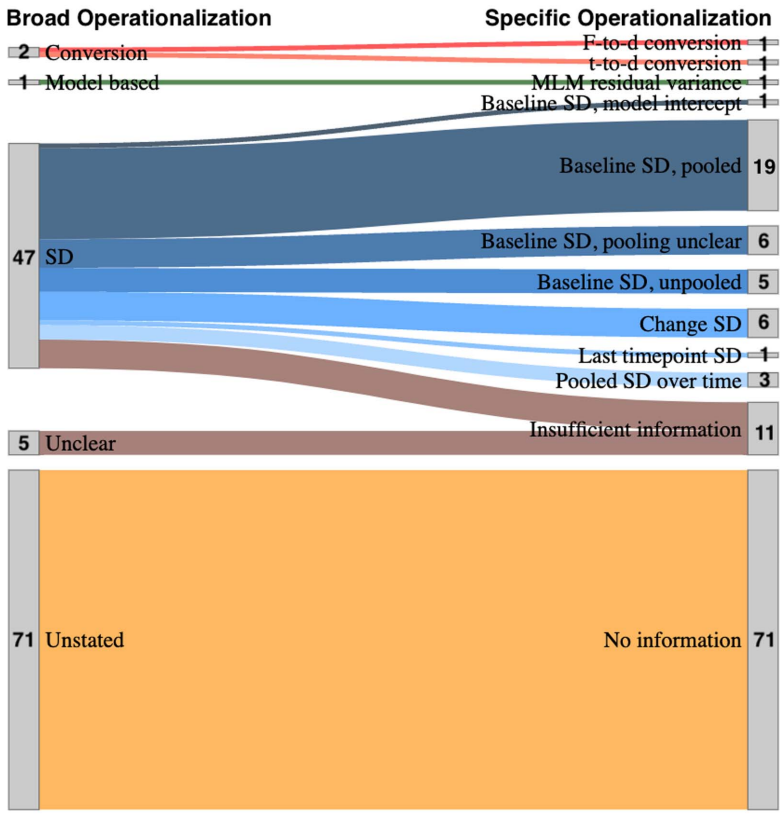
A brief summary of the review procedure is provided; a detailed description can be found at Crenshaw et al. (2025). Ten undergraduate and graduate student research assistants, plus the author, downloaded all articles from the relevant years and journals, coded each with respect to whether it was a clinical trial, and then, for clinical trials, coded whether an effect size was reported. The author subsequently read each method and results section to determine whether the effect size included an SMD and coded its specific operationalization. All coders underwent 9 weeks of training until reaching 90% agreement with the author, a clinical and quantitative psychologist experienced in carrying out and testing clinical trials in psychology and conducting research on trial methodology. Weekly coders' meetings were held to address questions. The data and coding manual can be found at https://doi.org/10.17605/OSF.IO/CV83U.

Of 626 articles coded, 226 (36%) were clinical trials. Among clinical trials, 126 (56%) reported at least one SMD. Figure 1 shows the SMD operationalizations according to the broad operationalization (the stated or cited method) and specific operationalization (how it was implemented). Only 44 (35%) studies provided sufficient information to determine the SMD operationalization, with 71 (56%) reporting no information and 11 (9%) reporting insufficient information. Of these 44 studies, we identified at least nine ways in which the SMD was operationalized, including dividing MD estimates by a baseline *SD* (pooled: 19, 43%; unpooled: 5, 11%; pooling unclear: 6, 14%; baseline *SD* using model intercept: 1, 2%), change *SD* or equivalent (6, 14%), *SD* pooled across time (3, 7%), and last time point *SD* (1, 2%). Three methods (7%) computed the SMD from model parameters.

[1] Whereas each specific term originally reflected a specific operationalization of the SMD, usage of terms like "Cohen's *d*" in modern times has shifted to simply indicate an SMD, where the specific operationalization is only known if described (Cumming, 2013; Luo et al., 2022).

Figure 1

Broad (Left Side) and Specific (Right Side) Operationalization of the Standardized Mean Difference in Reviewed Clinical Trials



Note. Broad Operationalization = general approach or method cited; Specific Operationalization = the specifics of how the method was implemented. "Pooled" *SD* refers to computation of one *SD* across groups, whether by computing the *SD* for the whole sample or by using a formula to pool two separate *SD* estimates. MLM = multilevel model. See the online article for the color version of this figure.

These results indicate that operationalization of the SMD varies widely across published clinical trials in psychology.

Toward Greater Standardization of the SMD

The proliferation of different operationalizations for the standardizer undermines the purpose of standardization so central to the SMD. Although this problem may be intractable in other research domains due to the sheer number of different study designs and purposes, many clinical trials in psychology share similarities in both design and purpose. Trials typically assess outcome variables before treatment and at one or more posttreatment time points. The primary estimand of interest is generally the between-arm difference at one or more posttreatment time points (for RCTs)[2] or change from baseline (for uncontrolled trials). Finally, for RCTs, patients are randomly assigned to two or more conditions.

Can we adopt a one-size-fits-most standard operationalization of the SMD's standardizer for clinical trials? If it were a simple feat, this issue would already be solved. Indeed, many scholars have made recommendations toward this end, typically listing one to a few favorable properties of the recommended approach or downsides of other approaches (e.g., Cumming, 2013; Daly et al., 2021; Downing et al., 2025; Feingold, 2013; Hopkins & Rowlands, 2024; Ostinelli et al., 2024). However, all approaches have limitations, and there has not been a comprehensive evaluation of the pros and cons of each approach.

In the remainder of this commentary, I will describe desired properties of the SMD and evaluate each standardizer with respect to these properties. Evaluation is done with respect to two goals: (a) determine whether clearly inferior options can be eliminated to achieve an operational de-proliferation of sorts and (b) determine whether it is possible to arrive at a default standard operationalization, recognizing that there will always be exceptions. I will focus primarily on RCTs but mention considerations for uncontrolled trials as well. Table 1 lists each property and notes whether each short-listed operationalization

[2] The baseline-adjusted between-condition difference at posttreatment (i.e., ANCOVA model; common in medical trials), unadjusted difference, and difference in pre–post change (common in psychology trials) all provide unbiased estimates of the same estimand in a properly randomized RCT (e.g., Higgins et al., 2024; Feingold, 2013). However, the ANCOVA model is more efficient (e.g., Daly et al., 2021; Higgins et al., 2024).

Table 1
Desirable Properties of the SMD and Extent to Which Each Standardizer Meets Each Property

Characteristic	Standardization method					
	Baseline *SD* (pooled)	Post-*SD* (pooled)	Post-*SD* (control group)	Change *SD* (pooled)	Baseline + post-*SD* (pooled ×2)	Model-derived
1. Sign agreement with raw mean difference	✓	✓	✓	✓	✓	✓
2. Transparent population	✓		✓	X		X
3. Population is comparable across studies	✓		✓–			
4. Minimize SMD inflation from inclusion criteria		✓	✓	✓	✓–	✓
5. Precision	✓	✓		✓–	✓✓	✓
6. Equal variance assured or not needed	✓		✓			Depends on method
7. Unaffected by loss to follow-up	✓					
8. Applicable to most trial designs	✓					

Note. Pooled ×2 = pooled across condition and time; ✓ = meets property well; ✓✓ = meets property very well; ✓– = somewhat meets property or meets in only some situations; X = badly violates property; SMD = standardized mean difference.

(after Property 1 eliminations) meets the property. The Open Science Framework page includes a table evaluating all considered operationalizations (Crenshaw, 2026).

Property 1 (Required): Sign Agreement With the Raw MD

The operationalization should not lead to nonsensical results, such as situations where one group is superior on the MD but inferior on the SMD. Use of group-specific (treatment arm, subgroup, etc.) *SD*s, where each group is standardized on its own *SD* before being compared, can lead to this outcome. For example, imagine a quality of life outcome where Condition A has a mean pre–post change of 10 and arm-specific *SD* of 12 and Condition B has a mean pre–post change of 8 and arm-specific *SD* of 8 (alternatively, substitute "change" with "[baseline-adjusted] post-treatment score"). In this case, the MD would favor Condition A ($10 - 8 = 2$), but the SMD using an arm-specific *SD* would favor Condition B ($\frac{10}{12} - \frac{8}{8} = -0.17$).

A properly operationalized SMD cannot be one that conflicts with the MD on something as fundamental as the direction of difference. Therefore, group-specific *SD*s typically should not be used. Six options remain for the choice of standardizer: baseline *SD* pooled across conditions ($SD_{\text{bl-pooled}}$), posttreatment *SD* pooled across conditions ($SD_{\text{postpooled}}$), posttreatment control group *SD* ($SD_{\text{post-c}}$), *SD* pooled across conditions and across time points (SD_{pooled2}), *SD* of change scores pooled across conditions ($SD_{\text{change-pooled}}$), and a model-derived *SD* (SD_{model}).

Property 2: Population of Comparison Is Transparent

When considering which value of the *SD* by which to divide estimates, one is effectively choosing the population distribution against which the raw estimates are compared (e.g., Cumming, 2013; Daly et al., 2021). Different populations may have different levels of variability, which impacts the *SD* and therefore SMD. For example, samples with restricted range (e.g., clinical samples) will tend to have less variability than those without (e.g., community samples). Consequently, the population should be transparent so readers and meta-analysts can appropriately contextualize the magnitude of the SMD.

Both SD_{model} and $SD_{\text{change-pooled}}$ suffer here because the comparison population is opaque under the best of circumstances (Feingold, 2013; Hopkins & Rowlands, 2024)—for example, it is much more difficult to conceive of a distribution of change scores or of model-derived estimates than of a distribution of raw scores. $SD_{\text{bl-pooled}}$ has several advantages. Variability in postbaseline time points is influenced by the specific treatment, with some treatments producing more heterogeneous outcomes than others (Downing et al., 2025; Hopkins & Rowlands, 2024). Yet most readers are unlikely to intuit how such heterogeneity impacts the distribution of posttreatment scores to properly contextualize the population of comparison. The $SD_{\text{bl-pooled}}$ is naturally free of treatment influences. Moreover, a natural comparison population for estimates of a treatment effect is the population that is eligible for the treatment but has not (yet) begun treatment—that is, the group at baseline (Cumming, 2013; Daly et al., 2021). $SD_{\text{post-c}}$ also does well in that a control group population is a natural comparison and is easy to conceptualize (e.g., Hopkins & Rowlands, 2024). $SD_{\text{postpooled}}$ and SD_{pooled2} fall somewhere in the middle.

Property 3: Population Is Comparable Across Studies

Populations will differ across studies to some extent, often greatly, but the SMD operationalization should enable *comparison* of the populations and should ideally not exacerbate differences in populations. $SD_{\text{change-pooled}}$ has very poor comparability because it experiences large swings depending on the magnitude of the pre–post correlation in the particular study (Harrer et al., 2025; Higgins et al., 2024; Hopkins & Rowlands, 2024) and is likely influenced by idiosyncratic study characteristics like the length of time between measurements (Higgins et al., 2024). SD_{model} suffers from these same problems and more because these methods are often conceptually similar to $SD_{\text{change-pooled}}$ (Feingold, 2013), and there is almost no consistency across studies in the particular method used for model-derived *SD*s (Luo et al., 2022).

$SD_{postpooled}$ and $SD_{pooled2}$ are also not ideal because the population of comparison differs depending on the heterogeneity of the particular treatment effects (e.g., Hopkins & Rowlands, 2024). $SD_{post\text{-}c}$ is more comparable because outcomes in a control group will not be impacted by the particular active treatment being studied. However, there are still many different possible control group conditions (e.g., waitlist, treatment as usual, pill placebo), which provide different populations for comparison for studies using different control groups.

$SD_{bl\text{-}pooled}$ meets this property well, though it is not perfect. Populations will always differ across studies based on design characteristics like inclusion criteria and sampling strategy. However, it best meets this principle because it is not impacted by the particular treatment, model, or comparison condition. Additionally, use of $SD_{bl\text{-}pooled}$ is conceptually most equivalent to SMDs computed in cross-sectional designs (Feingold, 2013), allowing the meaning of the SMD to stay relatively similar even across large study design differences.

Property 4: Minimize SMD Inflation Caused by Range Restriction From Inclusion Criteria

Clinical trials commonly employ inclusion criteria that screen out low-severity individuals (Harrer et al., 2025), creating a restricted range on the outcome variable and potentially inflated SMD estimates (Furukawa & Leucht, 2013). Because it is most directly impacted by the restricted range, $SD_{bl\text{-}pooled}$ suffers here in studies employing severity criteria. Some evidence suggests this issue is less pronounced in postbaseline time points. Furukawa and Leucht (2013) provided the most direct evidence: When the hypothetical severity criterion was raised in three schizophrenia medication trials, all other factors held constant, baseline *SD*s became smaller, and posttreatment *SD*s increased or increased slightly. Results from other studies are largely consistent with this explanation (Gallardo-Gómez et al., 2024; Harrer et al., 2025; Luo et al., 2022; Ostinelli et al., 2024), though it is difficult to isolate the effects of severity criteria from other potential causes like treatment heterogeneity effects on posttreatment *SD*s. $SD_{postpooled}$ and $SD_{post\text{-}c}$ are favored here, with $SD_{pooled2}$ somewhere in the middle.

Clinical and community populations are legitimately different populations, so the resultant range restriction caused by reasonable severity criteria is not necessarily problematic. A larger problem would be if companies, app developers, device manufacturers, treatment developers, or other trialists with treatment allegiances attempted to deliberately restrict the baseline *SD*—such as by employing overly restrictive severity criteria or through recruitment strategies—to inflate the SMD, which can then impact meta-analytic estimates and ultimately treatment recommendations. In such situations, $SD_{bl\text{-}pooled}$ would be most susceptible, with evidence suggesting other standardizers would be less so. However, such manipulation would also carry significant costs. A trial's inclusion criteria face evaluation by collaborators, grant reviewers, ethics boards, and journal reviewers. Severity criteria that are not justifiable on scientific, ethical, and feasibility grounds would be at higher risk of rejection at these stages, and extreme examples are likely to fail outright. Additionally, employing overly high severity criteria limits the recruitment pool in a context in which meeting recruitment targets is already a perpetual worry among trialists. Therefore, while manipulating design elements to artificially inflate SMD estimates would be a significant problem and may occur in isolated contexts, it is unlikely to be a widespread practice even among bad-faith actors due to its associated costs.

Property 5: Precision

Sampling error creates noise in estimates, including the estimate of the *SD* (Downing et al., 2025). Therefore, approaches that minimize sampling error are desirable. This principle favors pooling *SD*s across conditions and/or time points to estimate with a larger sample compared with unpooled *SD*s (Feingold, 2013). All approaches have an advantage here except $SD_{post\text{-}c}$, with an extra benefit of $SD_{pooled2}$ due to pooling across both groups and time points.

Property 6: Ensures Equal Variance Assumption Is Met or Does Not Require It

Pooling *SD*s between groups or measurement occasions rests on the equal variance assumption, where any difference in *SD*s is due only to sampling error and not to genuine differences between groups (e.g., Feingold, 2013). Therefore, it is desirable either to assure equal variance is met or to not pool. $SD_{bl\text{-}pooled}$ satisfies the equal variance assumption in a properly randomized trial because randomization ensures population equivalence at baseline. $SD_{post\text{-}c}$ avoids the issue by not pooling. In contrast, equal variance would need to be tested for $SD_{postpooled}$ and $SD_{pooled2}$, and it is not clear what should be done when the assumption is violated. $SD_{pooled2}$ doubly suffers by assuming equal variance across both condition and time.

Property 7: Standardizer Unaffected by Loss to Follow-Up

Loss to follow-up can bias results when those who drop out from study assessments are systematically different from those who do not. Therefore, any measure occurring after baseline has the potential to be biased by loss to follow-up, particularly if the loss is a large proportion of participants. $SD_{bl\text{-}pooled}$ is not exposed to this risk. $SD_{pooled2}$ does slightly better than the remaining methods as it combines measurement occasions that are and are not exposed to this risk.

Property 8: The Operationalization Can Be Consistently Applied to Most Clinical Trial Designs

It is desirable for the chosen method to be usable across many common clinical trial designs so that SMDs have a common meaning and are maximally comparable to one another. Therefore, methods that cannot be used or face additional complications for other common study designs beyond the prototypical pre–post between-group comparison trial are disfavored. $SD_{bl\text{-}pooled}$ meets this property well because collecting a baseline assessment is standard practice in psychology clinical trials. Other operationalizations suffer. For posttreatment *SD*s, it is not clear what to do for studies with multiple posttreatment assessments (e.g., posttreatment and follow-up, crossover designs; Hopkins & Rowlands, 2024). Use of the time-specific *SD* can again lead to dissociations of MD and SMD results. $SD_{post\text{-}c}$ suffers an additional downside because it cannot be used for uncontrolled trials or RCTs that compare only active treatments.

Additionally, $SD_{\text{change-pooled}}$ is questionable when the MD is not a change estimate (e.g., for analysis of covariance [ANCOVA] models).

A One-Size-Fits-Most Approach

The pooled baseline *SD* meets all but one of the desired properties of an SMD, far more than any other standardizer: It avoids nonsensical results like sign or magnitude reversal when comparing groups or time points, is relatively precise, ensures the equal variance assumption is met, is unaffected by loss to follow-up, is usable, and can be consistently applied for most psychology clinical trial designs, and it provides a transparent population that is maximally comparable across studies. Its main flaw is susceptibility to range restriction, including the possibility of deliberate manipulation discussed under Property 4. Not all properties are weighed equally, and given the potential costs, this property is weighed heavily.

However, several factors temper this concern. The natural disincentives to manipulating study design elements to inflate SMD estimates are significant. The limitation is also readily apparent to readers of individual trials because describing inclusion criteria is standard reporting, though it would be less visible in meta-analyses. Additionally, future work that systematically estimates the influence of different severity criteria on baseline *SD*s might lead to a baseline *SD* correction for such studies. Many limitations of other methods are not so transparent, so readers are left guessing as to the implications of the limitations, or they may not be aware of them at all.

Reasonable scholars may weigh the properties in Table 1 differently, and the reporting recommendations in the next section enable use of alternative standardizers when warranted. However, the risk of SMD inflation in Property 4 would have to be exceptionally large to overcome the breadth of advantages for $SD_{\text{bl-pooled}}$. Therefore, $SD_{\text{bl-pooled}}$ provides the strongest available basis for a default recommendation.

Two additional objections to use of the baseline *SD* have been raised. Some sources argue that the posttreatment *SD* is more appropriate if baseline is not involved in the MD estimate, such as in ANCOVA models (e.g., Furukawa & Leucht, 2013). However, while the *SD* should certainly be *applicable* to the MD (e.g., dividing a MD in depression by a *SD* in grade point average would be incoherent), it is not clear why it would need to come from the exact distribution. Indeed, there have been many calls for meta-analytic estimates to be standardized on an external reference *SD* as opposed to each sample's *SD* (Daly et al., 2021; Downing et al., 2025; Gallardo-Gómez et al., 2024). The SMD is simply a unit conversion, a reexpression of the MD estimate under a chosen reference distribution. As long as the chosen unit is coherent, the *SD* does not need to be derived from the exact distribution as the MD estimate.

Second, the extent to which a treatment produces heterogeneity in response is an important feature of a treatment. One might argue that, therefore, this heterogeneity should be reflected in the SMD, such as by standardizing on $SD_{\text{postpooled}}$. However, this argument blends two constructs that are best kept distinct. The MD is a mean estimate. Standardizing the MD with an *SD* free of the influence of treatment effect heterogeneity maintains its status as a mean estimate. In contrast, treatment effect heterogeneity is a form of variability and is better evaluated by assessing variability directly, such as random slope analysis, idiographic-based methods like number of treatment responders, or moderator or subgroup analysis.

One alternate method increasingly discussed in the meta-analytic literature is standardizing the MD with an external reference *SD* from a representative study (e.g., Daly et al., 2021; Downing et al., 2025). However, establishing this practice for individual trials would require consensus on a system for selecting appropriate reference values to prevent SMDs from again becoming incomparable. It would also be necessary to address the potential to inflate SMD estimates by selecting the smallest justifiable reference *SD*, which would be far easier and face fewer disincentives than manipulating design elements. Therefore, while intriguing, much more work is needed before it can be recommended as a general approach.

Reporting Recommendations

Echoing prior calls (e.g., Cumming, 2013; Daly et al., 2021; Harrer et al., 2025), better and more consistent reporting of how SMDs were operationalized is urgently needed. Only one third of psychology clinical trials that computed an SMD reported sufficient information to understand its operationalization, closely matching the 31% reported for medical trials (Luo et al., 2022). Given the large potential impact of different operationalizations on the magnitude of SMD estimates demonstrated by prior investigations, improving reporting is low-hanging fruit for improving practices in usage of the SMD. I recommend the following:

1. Report how the MD was computed so the numerator of the SMD formula is clear. This will depend on the model used and may be, for example, the baseline-adjusted between-group difference at posttreatment, the model-estimated between-group difference in pre–post change, or (for uncontrolled trials) the model-estimated pre–post change.

2. Report the specific *SD* used in the denominator, including any pooling conducted—for example, "pooled baseline *SD*." Merely citing a specific article is insufficient, no matter how seemingly clear, as my review found numerous incidents of researchers using different operationalizations while citing the same article (e.g., Feingold, 2013).

3. Use the term "SMD" in place of terms like Cohen's *d*, Hedges's *g*, or Glass's Δ. The meaning of these terms has become muddied over time (Cumming, 2013). Therefore, it is preferred to use the generic "SMD" and describe the specific operationalization. However, small sample corrections can be appropriately described as, for example, a "Hedges's *g* correction" (Hedges, 1981).

4. Prespecify the specific SMD operationalization in advance given the impact of the operationalization on the magnitude of the SMD estimate (Harrer et al., 2025).

5. Always report the MD estimate alongside SMD estimates. Doing so allows other readers and meta-analysts to compare results more precisely, which is best done in original units when studies use the same measure (Daly et al., 2021; Gallardo-Gómez et al., 2024).

6. Report means and *SD*s for all major study assessment time points, which allows other users and meta-analysts to recompute SMDs using a different *SD* if desired. This practice can also show whether the baseline *SD* may have been artificially reduced, which can serve as a safeguard against deliberate SMD inflation.

Limitations and Exceptions

Use of the pooled baseline *SD* is intended as a one-size-fits-most approach to cover most common clinical trial designs, not a rigid prescription for all situations. There will be exceptions in which it cannot be used or another approach is preferred. These include trials without baseline assessments (e.g., posttest-only RCTs), outcomes that only exist after treatment begins (e.g., treatment satisfaction), studies in which the estimand is not a MD (e.g., if variability is of principal interest), and estimands focused on a different level of analysis (e.g., cluster-level outcomes; Feingold, 2013). Additionally, the SMD is only appropriate for continuous outcomes.

Exceptions also include trials in which baseline and follow-up populations are qualitatively different. Participants in long-term developmental data sets age into different populations, or they may be measured on outcomes that can only be assessed after a certain level of development (e.g., antisocial personality). Subgroup analyses conditioned on a later outcome (e.g., assessing the proportion of treatment responders who later relapse) change the population from baseline, so standardizing with the concurrent time point likely makes more sense. Additionally, an intervening event may fundamentally alter the population from baseline (e.g., emergency room health care workers measured both before and after the COVID-19 pandemic). Nonrandomized trials and subgroup analyses cannot ensure equal variance for pooling, though it is still usually preferable to put scores from all groups on the same unit.

Conclusion

The SMD is an important tool for evaluating treatments in psychology clinical trials. It enables consumers of research to understand magnitudes without knowing the underlying distribution of every outcome, provides a common language for researchers and clinicians alike to communicate, and enables quantitative aggregation of studies in meta-analyses. As the name suggests, its utility is maximized when its definition does not change across users. Given commonalities in purpose and design of clinical trials that enable usage of a common operationalization, it is time to standardize the computation of the SMD.

References

Crenshaw, A. O. (2026, April 1). *Evaluating change in psychology clinical trials*. https://doi.org/10.17605/OSF.IO/CV83U

Crenshaw, A. O., MacDonald, A., Carr, K., Shull, A., Pope, S., & Blackmon, M. (2025). *Inconsistent application of the reliable change index in clinical psychology trials leads to noisy estimates of individual change*. https://doi.org/10.31234/osf.io/d9amr_v1

Cumming, G. (2013). Cohen's d needs to be readily interpretable: Comment on Shieh (2013). *Behavior Research Methods*, *45*(4), 968–971. https://doi.org/10.3758/s13428-013-0392-4

Daly, C., Welton, N. J., Dias, S., Anwer, S., & Ades, A. (2021). *Meta-analysis of continuous outcomes: Guideline methodology document 2. NICE guidelines technical support unit*. https://www.bristol.ac.uk/population-health-sciences/centres/cresyda/mpes/nice/guideline-methodology-documents-gmds/

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*, *16*(3), 460–478. https://doi.org/10.1017/rsm.2025.7

Feingold, A. (2013). A regression framework for effect size assessments in longitudinal modeling of group differences. *Review of General Psychology*, *17*(1), 111–121. https://doi.org/10.1037/a0030048

Furukawa, T. A., & Leucht, S. (2013). Can we inflate effect size and thus increase chances of producing "positive" results if we raise the baseline threshold in schizophrenia trials? *Schizophrenia Research*, *144*(1–3), 105–108. https://doi.org/10.1016/j.schres.2012.12.006

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*, Article 111213. https://doi.org/10.1016/j.jclinepi.2023.11.003

Harrer, M., Miguel, C., Luo, Y., Ostinelli, E. G., Karyotaki, E., Leucht, S., Furukawa, T. A., & Cuijpers, P. (2025). Standardized effect sizes are far from "Standardized": A primer and empirical illustration in depression psychotherapy meta-analyses. *PLOS Mental Health*, *2*(7), Article e0000347. https://doi.org/10.1371/journal.pmen.0000347

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane handbook for systematic reviews of interventions* (Version 6.5). Cochrane. https://www.cochrane.org/handbook

Hopkins, W. G., & Rowlands, D. S. (2024). Standardization and other approaches to meta-analyze differences in means. *Statistics in Medicine*, *43*(16), 3092–3108. https://doi.org/10.1002/sim.10114

Luo, Y., Funada, S., Yoshida, K., Noma, H., Sahker, E., & Furukawa, T. A. (2022). Large variation existed in standardized mean difference estimates using different calculation methods in clinical trials. *Journal of Clinical Epidemiology*, *149*, 89–97. https://doi.org/10.1016/j.jclinepi.2022.05.023

Ostinelli, E. G., Efthimiou, O., Luo, Y., Miguel, C., Karyotaki, E., Cuijpers, P., Furukawa, T. A., Salanti, G., & Cipriani, A. (2024). Combining endpoint and change data did not affect the summary standardised mean difference in pairwise and network meta-analyses: An empirical study in depression. *Research Synthesis Methods*, *15*(5), 758–768. https://doi.org/10.1002/jrsm.1719

Sahker, E., Luo, Y., Omae, K., Tajika, A., Ferreira, M. L., Chevance, A., Leucht, S., Markham, S., Ede, R., Cipriani, A., Salanti, G., & Furukawa, T. A. (2025). Estimating the smallest worthwhile difference of recommended psychotherapies for depression: Observational study. *The British Journal of Psychiatry*. Advance online publication. https://doi.org/10.1192/bjp.2025.10453

Received September 5, 2025
Revision received March 25, 2026
Accepted March 30, 2026 ■