Draft Methods Research

Number XX

Effect Size Metric Choices as Factors in Meta-Analytic Statistical Inferences

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

This information is distributed solely for the purposes of pre-dissemination peer review. It has not been formally disseminated by the Agency for Healthcare Research and Quality. It does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services (AHRQ) determination or policy.

AHRQ Publication No. xx-EHCxxx
<Month Year>

This report is based on research conducted by the XXXXX Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. XXX-XXX-XXXXX). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decision makers—patients and clinicians, health system leaders, and policymakers, among others—by improving the methods that meta-analyses use to accumulate data about health care services and other matters. This report is not intended to apply clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials that are clearly noted in the document. Further reproduction of those copyrighted materials is prohibited without the specific permission of copyright holders.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact [insert program e-mail address].

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: To be added in final version of report.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Parivash Nourjah
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

Peer Reviewers

<Name>
<Place>
<City>, <ST>

<Name>
<Place>
<City>, <ST>

Effect Size Metric Choices as Factors in Meta-Analytic Statistical Inferences

Structured Abstract

Introduction. Meta-analysis cannot proceed unless each study outcome is on the same metric and has an appropriate sampling variance estimate, the inverse of which is used as the weight in meta-analytic statistics. When comparing treatments for trials that use the same measures across studies, contemporary meta-analytic practice uses the original (raw) unstandardized mean difference (UMD) to model the difference between the observed means (i.e., M_E-M_C) rather than representing effects in the standardized mean difference (SMD). A fundamental difference between the two strategies is that the UMD incorporates the observed variance of the measures as a component of the analytical weights (viz., sampling error or inverse variance) in statistically modeling the results for each study. In contrast, the SMD incorporates the measure variance directly in the effect size (ES) itself (i.e., SMD=[M_E-M_C]/SD) and not directly in the analytical weights. The UMD approach has been conventional despite the fact that its bias and efficiency are unknown and have not been compared to the SMD. Also unresolved is which of many possible available equations best optimize statistical modeling for the UMD and SMD (and versions of these equations that examine repeated measures versus between-groups (or mixed) designs).

Estimates of SMD and its variance depend on whether the comparison is derived from a repeated measures or between groups design and on the particular statistical approach used and they are not always equivalent. Although several methods exist to estimate SMD and its variance, none provide an unbiased estimation and the circumstances under which they are most comparable is currently unknown. Three designs are compared: posttest-measures two-groups, repeated-measures one-group, and repeated-measures two-groups design. The SMD for the last two can be estimated using a change-score metric or a raw-score metric of the effect size, depending whether the correlation between pretest and posttest is used to standardize the effect.

Methods. Monte Carlo simulations compared available equations in terms of their bias and efficiency across the many different conditions established by crossing: (1) number of studies in the meta-analysis (k = 10, 20, 50, and 100); (2) mean study sample sizes (5 values of N ranging from small to very large); (3) the ratio of the within-study observed measure variances for experimental and control groups and at pretest and posttest (ratios: 1:1, 2:1, and 4:1); (4) the posttest mean of each pseudo experimental group to achieve 3 parametric effect sizes (μ^{E}_{post} = 0.25, 0.50, and 0.80); (5) normal versus non-normal distributions (4 levels); (6) the between-studies variance (τ^2= 0, 0.04, 0.08, 0.16, and 0.32). For the second issue, (7) the correlation between the two conditions was manipulated through the variance-covariance matrix, where homogeneous variances equal to 1 can be assumed. The manipulated correlation was equal to the covariance between the two measures ($\rho_{pre\text{-}post}$ = 0, 0.25, 0.50, and 0.75).

Results. Simulations showed that leaving the effect size (ES) index in the original metric presents no bias or loss of efficiency when distributions are normal, when there is no

heterogeneity among effects, and when the observed variances of the experimental and control groups are equal; in contrast, when conditions deviate from these ideals, standardizing outcomes is markedly better. When there is high skewness and kurtosis, neither metric has a marked advantage. Thus, the standardized index presents the least bias under most conditions and is more efficient than the unstandardized index. The unstandardized mean difference ES has higher efficiency only with very small within-study variances or study sample sizes.

Standardized effect sizes can be recommended under all conditions. It is particularly the case that the standardized mean difference will perform better when within-study variability is large, when parametric assumptions are poorly met, and when sample sizes increase. Finally, the results comparing all the SMD and its variance estimations show both the utility of a new solution and conditions under which different estimation procedures may affect results depending on the design and the metric. All estimations reach similar inferences when they do not include the correlation between pretest and posttest or if the correlation between measures is 0.50. The current results imply that the choice of effect size metrics, estimators, and sampling variances can have large effects on statistical inferences even under such commonly observed circumstances as normal sampling distributions, large samples of studies and studies with large samples, and when effects exhibit heterogeneity. Thus, this investigation provides guidance for improved statistical practice in relation to meta-analysis of literatures comprised of studies that compare two groups at one point in time, or that examine repeated measures for one or two groups. The Discussion considers clinical interpretation of results using the SMD and addresses limitations of the current project.

Contents

Introduction **1**
Background 1
Objectives 2
Project Specific Terminology 4
Methods **12**
General Design 12
Results **14**
Part One: Unstandardized Versus Standardized Effect Sizes 14
Overview 14
Detailed Analysis 14
Part Two: Using the SMD with Repeated Versus Independent Measures 19
Overview 19
Detailed Analysis 20
Discussion **26**
References **30**

Tables

Table 1. Standardized mean difference ES estimations (and their components) for a one-group repeated-measures design 7
Table 2. Standardized mean difference (SMD) ES estimations (and their components) for two independent groups. 8
Table 3. Estimates of sampling variance for the SMD ES in the one-group design with repeated-measures 10
Table 4. Estimates of sampling variances for the SMD ES from two-group designs with repeated-measures 10
Table 5. Estimates of statistics related to the unstandardized mean difference (UMD) for designs with two independent groups 11
Table 6. Characteristics of Simulated Datasets 14
Table 7. Recommendations for Effect Size Metric and Relevant Formulas 27

Figures

Figure 1. Bias of SMD and UMD as a function of asymmetry and the between-studies variance of the distribution. 15
Figure 2. Bias of SMD and UMD as a function of the between-studies variance of the distribution. 15
Figure 3. Bias of SMD and UMD as a function of the parametric effect size. 16
Figure 4. Bias of SMD and UMD as a function of mean sample size. 16
Figure 5. Bias of SMD and UMD as a function of heteroskedasticity. 16

Figure 6. Efficiency of SMD and UMD as a function of asymmetry. 17
Figure 7. Efficiency of SMD and UMD as a function between-studies variance. 17
Figure 8. Efficiency of SMD and UMD as a function of the parametric effect size. 18
Figure 9. Efficiency of SMD and UMD as a function of mean sample size. 18
Figure 10. Efficiency of SMD and UMD as a function of heteroskedasticity. 18
Figure 11. Bias as a function of the correlation parameter of the ES indexes. 20
Figure 12. The efficiency as a function of the correlation parameter of the ES indexes 21
Figure 13. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes. 22
Figure 14. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes. 22
Figure 15. The bias as a function of the correlation parameter of the ES indexes. 23
Figure 16. Efficiency as a function of the correlation parameter of the ES indexes. 24
Figure 17. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes. 24
Figure 18. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes. 25

Appendixes

Appendix A. Bias and Efficiency Results for Standardized and Unstandardized Mean Differences
Appendix B. Bias, Efficiency, and Theoretical Variance for All Effect Size Indexes and Their Variances.

Introduction

Background

Over the past 30 years, meta-analytic methods to accumulate knowledge have experienced a sharp increase in use across the sciences, and have been applied to many topics of high import to public health. Using meta-analysis, the result of every study is quantified by means of a statistical index that one can apply to all studies in a given literature, thereby enabling a comprehensive summary of the magnitude of the effect in every study and analyses of outcomes according to coded study features.[1-11] Conventionally, meta-analysis has three main objectives: (1) synthesizing different studies' effect size values to obtain a weighted mean, (2) assessing the consistency of the results, and (3) in the case of inconsistency (or heterogeneity), using moderator variables in an attempt to explain the variability. To do their work, meta-analysts must complete a series of interrelated steps: (1) conceptually define the topic of the review, (2) set selection criteria for the sample of studies, (3) comprehensively search for qualified studies, (4) code studies for their distinctive substantive, methodological and external characteristics, (5) represent the magnitude of each study's effect on the same metric, (6) analyze the database, and (7) interpret and present the results. To the extent that meta-analysts have the best available techniques to complete each step, the accuracy of their conclusions will be enhanced; science and its applications can accumulate and report its research findings in a more efficient manner.

The current report focuses on the 5th and 6th steps: Meta-analysis cannot proceed unless each study outcome is represented on the same metric and has an appropriate sampling variance estimate, the inverse of which is used as the weight in meta-regression and other meta-analytic statistics. In contemporary practice, when comparing treatments for trials that use the same measures across studies, meta-analyses routinely use the original or unstandardized mean difference (UMD) to model the difference between the observed means (i.e., M_E-M_C) rather than representing effects in the standardized mean difference (SMD). A fundamental difference between the two strategies is that the UMD incorporates the observed variance of the measures as a component of the analytical weights (viz., sampling error or inverse variance) in statistically modeling the results for each study. In contrast, the SMD incorporates the measure variance directly in the effect size (ES) itself (i.e., SMD=[M_E-M_C]/SD) and not directly in the analytical weights. The UMD approach has been conventional despite the fact that its bias and efficiency are unknown and have not been compared to the SMD. Also unresolved is which of many possible available equations best optimize statistical modeling for the UMD and SMD. There are many equations that examine repeated measures and there are many potential estimates of the inverse variance needed for modeling either UMD or SMD.

Another important and controversial issue is specifically related to the SMD. This estimator is used to measure the degree of change between repeated measures or the difference between two groups, using a standardization that can vary depending on the standard deviation used, with the assumption that the measures follow a normal distribution. In its between-groups form, SMD can be calculated from any two groups whether they are experimental or not; it is assumed that the individuals in the compared groups are independent. In its repeated-measures or within-subjects form, SMD assumes that the observations are dependent, and while some extant meta-analytic procedures account for this dependency, many others do not; scholars will often integrate both types of estimates in a single meta-analysis. Similarly, the numerous methods of calculating the SMD and their variances are known to vary.[17-18]

Therefore, it is unknown how much bias appears in the weighted effect sizes and posterior analyses when two-groups, two-groups repeated measures, or single groups with repeated measures are integrated without incorporating assumptions about possible dependence arising from the those observations with repeated measures. Further, it is not clear in the literature if the different methods of transformation to standardized mean difference from different statistical information types are equal across design types. There is conflicting advice about which specific technique equations to invoke when trials assess an outcome on the same measure and/or evaluate outcomes using repeated measures versus between-groups (or mixed) designs.

Objectives

This report has two objectives:

1. Determine the bias and efficiency of the unstandardized mean difference (UMD) relative to the standardized mean difference (UMD) under a wide range of analytic circumstances.

In literatures for which a phenomenon is assessed using the same measure in every study, meta-analysts have the choice of examining either standardized effect sizes (*d*, *OR*, etc.) or leaving study outcomes in the original, unstandardized, measure.[12] For example, blood pressure is always assessed in millimeters of mercury (mm/Hg) and meta-analyses of blood pressure outcomes routinely leave it in this metric. Efficacy in anti-depressant trials is routinely assessed on the Hamilton Rating Scale of Depression (HAM-D), and many meta-analyses examine it in this metric. Analysts typically leave study results in the original unstandardized measure in order to facilitate their interpretability. Many prominent statisticians have even recommended leaving comparisons in a unstandardized measure to facilitate comparisons between studies.[13-15]

Such comparisons make a great deal of sense when the studies' samples exhibit equal variances on the measures of their phenomenon. Yet, to the extent that their studies have unequal measure variances, meta-analyses that leave results unstandardized may yield inferences that poorly represent the underlying studies outcomes. For example, anti-depressant trials focusing on very severely depressed individuals (e.g., *M* HAM-D=33) will have much larger standard deviations than trials that focus on moderately depressed individuals (e.g., *M* HAM-D=17). Making the problem worse, some trials select patients in narrower ranges of the scale, artificially making the underlying standard deviation smaller relative to other trials. Change of, say, 6 units on the HAM-D is more dramatic change for a sample with a small standard deviation than for a sample with a large one. In such instances, a standardized effect size metric (e.g., *d*) may solve the problem. Weights for unstandardized outcomes in meta-analysis routinely use the sample size and the measure variance,[12,16] but it is unclear whether this solution solves the problem. Homogeneity of compared group variances in primary-level research is an analogous assumption to the problem that appears in terms of between-studies heterogeneity in observed measurement variances.

The second objective is:

2. Determine the best techniques to calculate effect size estimates and their sampling variances for different designs and under different design and parametric conditions**.**

Accurate estimation of effect sizes and their sampling variances when studies with different designs are integrated. Meta-analysis cannot proceed unless each study outcome is on the same metric and an appropriate sampling variance can be calculated for it. Current meta-analytic methods yield conflicting advice about which specific techniques to invoke when the outcomes

are provided from different designs, specifically, within-, between-subjects, or mixed-design, again with the result that significance testing and interpretation may vary depending on how they are integrated.

An effect size estimator is used to measure the degree of change between repeated measures or to compare the difference between two or more groups, with the assumption that the measures follow a normal distribution. In its between-groups form, the effect size (ES) estimator can be calculated from any two groups whether they are experimental or not; it is assumed that the individuals in the compared groups are independent. In its repeated-measures or within-subjects form, the effect size estimator assumes that the observations are dependent, and while some meta-analytic procedures account for this dependency, many others do not; scholars often integrate both types of estimates in a single meta-analysis or they more simply focus on post-test results without incorporating baseline measures or the power of the different groups to detect the effect. Similarly, the numerous methods of calculating the ES when the outcome is continuous and their available variances are known to vary.[17,18] Therefore, it is unknown how much bias occurs in calculations of the weighted effect size and its posterior analysis in these different instantiations are integrated without assumptions about the possible dependence and other design and parametric conditions (e.g., sample size, number of studies, parametric effect size) arising from the two types of repeated measures observations. Further, the literature leaves unclear whether the different methods of transformation to standardized mean difference from different statistical information types are equivalent.[6,19-22]

Orientation to Method.

For both specific aims, statistical analytical work will be developed and Monte Carlo simulation studies will be used to test it generating data under a wide variety of conditions to determine the extent to which parameter estimates, sample sizes, and number of studies are unbiased and their standard errors efficient. The simulations will (a) evaluate the differences between using unstandardized versus standardized metric of effect size (objective 1); (b) evaluate current solutions to estimate the ES and its sampling variance, differentiating among three main design types (i.e., two-groups, two-groups repeated-measures, and repeated measures design) (objective 2); (c) develop new solutions as necessary (our team has generated some new alternatives for estimating repeated measures effect sizes and their sampling variances) (objective 2); and (d) conduct simulations under widely varying conditions to gauge the performance of these methods of estimation (objectives 1 and 2).

Significance of Project

The goals of this project are relevant to any empirical literature that has systematic observations; these concern statistical operations that are very commonly used in contemporary practice. Even if it turns out that meta-analytic statistics in the original metric are robust to underlying deviations in measure variance, the results of this investigation are of great interest. If meta-analytic statistics and inferences using unstandardized effect sizes *are* sensitive to deviations in unequal measure variance and normality assumptions (and other circumstances), then the findings have far-ranging implications for the practice of meta-analysis. Moreover, it is also important to know what estimates of within-subjects ESs (in single- and in two-group designs) are the most accurate, and to determine which estimates of variance, for use in conducting weighted analyses and when those two types of designs can be combined in a single meta-analytic database. Knowing how well each effect size index for each design performs will

enable future analysts a better choice of the most appropriate operations and, as a consequence, permit more studies to be integrated and more accurate meta-analytic results. Thus, this methodological study offers considerable information to improve the accuracy and progress of science and public health. An overarching goal is to permit scholars to incorporate greater numbers of studies in their meta-analyses and enable more accurate empirical generalizations.

Project Specific Terminology

Throughout this report we will use the following specific definitions:

- **Bias**: The extent to which the observed UMD or SMD differs from the parametric value. Positive values of bias imply over-estimations of the parametric effect size and negative values imply under-estimations.
- **Change-score metric**: The difference between two repeated measures compared to the variability of change scores
- **Effect size (ES):** In the current investigation, comparisons between groups, comparisons across time, or both, are effect sizes, or indexes of association.
- **Efficiency:** A measure of the optimality of an estimator that reaches the closest value to the parameter with the minimum variance through replicates To the extent that efficiency is positive, statistical power is maximized to detect the parametric value.
- **Mean Square Error (MSE)**: A measure of the average of the square of the errors that evaluates the quality of an estimator in terms of its variation and unbiasedness.
- **One-group repeated-measures design**: A study methodology in which a single sample is observed at two or more time points (e.g., before and after a treatment).
- **Raw-score metric**: A metric that compares a mean or mean difference between conditions or times with the variability of scores within each condition.
- **Sampling variance of effect size for one-group repeated-measures design**: The variance of the sampling distribution, which is the distribution of values that result from repeated random samples of the same size using a repeated-measures design (see Table 3 for extant estimations of this statistic).
- **Sampling variance of effect size for two independent groups:** The same as the preceding one but for studies following a two-groups design (see Table 4 for extant estimations of this statistic).
- **Standardized mean difference (SMD) effect size**: the difference of two means divided by the pooled standard deviation of underlying the two means.
- **Standardized mean difference (SMD) for one-group repeated-measures design**: The effect size comparing two means at different times for the same group relative to the standard deviation (see Table 1 for extant estimations of this statistic).
- **Standardized mean difference (SMD) for two independent groups:** The effect size the change in means for two independent groups (repeated-measures, between-groups version) or the comparison of the means at posttest for two independent groups (between-groups version). (See Table 2 for extant estimations of this statistic.)
- **Two-groups repeated measures design**: A study methodology in which two groups (or arms; e.g., treatment and control) are observed at two or more times.
- **Unstandardized mean difference (UMD) effect size**: the difference of the two means in their original metric or scale (where all studies use the same metric).

Following are the Greek terms that appear in this report:

Term	Definition
$\hat{\delta}_j$	The sample estimate of population parameter δ for the j_{th} replication
μ^E_{post}	Parametric mean at posttest for the experimental group
μ^E_{Pre}	Parametric mean at pretest for the experimental group
μ^C_{Post}	Parametric mean at posttest for the control group
μ^C_{Pre}	Parametric mean at pretest for the control group
σ^2_{Pre}	Parametric variance at pretest
σ^2_{Post}	Parametric variance at posttest
$\sigma_{Pre,Post}$	Covariance between pre- and posttest
ρ	Parametric correlation
$\rho_{pre\text{-}post}$	Parametric correlation between pre-and posttest
σ_C	Parametric standard deviation of the control group
σ_E	Parametric standard deviation of the experimental group
τ^2	Between-study variance

Following are the Latin abbreviations used in this report:

Term	Definition
d_b	Standardized mean difference proposed by Becker (See Table 1, No. 3 for details and elements of the equation)
d_{b_nonr}	Becker's standardized mean difference, which excludes the correlation factor $2(1 - r)$ in its equation
df	Degrees of freedom
d_g	Standardized mean difference proposed by Gibbons (See Table 1, No. 4 for details and elements of the equation)
d_{hb}	Standardized mean difference proposed by Hedges (See Table 2, No. 6 for details and elements of the equation)
d_{hw}	Standardized mean difference proposed by Huedo-Medina & Johnson (See Table 1, No. 5 for details and elements of the equation)
d_{hw_nonr}	Hedges' standardized mean difference using the within-study degrees of freedom and excluding the correlation factor $2(1 - r)$
d_{s1}	Standardized mean difference proposed by Shadish (See Table 2, No. 10 for details and elements of the equation)
d_{s2}	Standardized mean difference proposed by Shadish using the standard deviation from ANOVA results (See Table 2, No. 11 for details and elements of the equation)
d_{s3}	Standardized mean difference proposed by Shadish using the standard deviation from ANCOVA results (See Table 2, No. 12 for details and elements of the equation)
d_{tch}	Standardized mean difference based on *t*-test for change-score metric (See Table 1, No. 2 for details and elements of the equation)
d_{tra}	Standardized mean difference based on *t*-test for raw-score metric (See Table

Term	Definition
	1, No. 1 for details and elements of the equation)
ES	Effect size
HAM-D	Hamilton rating scale of depression
k	Number of studies
M HAM-D	Mean score on the HAM-D
M_C	Mean control
M_E	Mean experimental
mmHg	Millimeters of mercury (used in measures of blood pressure)
MSE	Mean square error
N	Total sample size
n	Group sample size
OR	Odds ratio
r	Estimated correlation
Rns	The number of replications
SD	Standard Deviation
SMD	Standardized mean difference (d)
UMD	Unstandardized mean difference
$var_{one\text{-}g}$ (d_b)	Variance estimate for one group design with repeated measures of the standardized mean difference proposed by Becker (See Table 3, No. 13 for details and elements of the equation)
$var_{one\text{-}g}$ (d_g)	Variance estimate for one group design with repeated measures of the standardized mean difference proposed by Gibbons (See Table 3, No. 14 for details and elements of the equation)
$var_{two\text{-}g}$ (d_b)	Variance estimate for two group design with repeated measures of the standardized mean difference as a function of two effect sizes proposed by Becker (See Table 4, No. 16 for details and elements of the equation)
$var_{two\text{-}g}$ (d_{b_t})	Variance estimate for two group design with repeated measures of the standardized mean difference proposed by Hedges (See Table 4, No. 15 for details and elements of the equation)
$var_{two\text{-}g}$ (d_g)	Variance estimate for two group design with repeated measures of the standardized mean difference as a function of two effect sizes proposed by Gibbons (See Table 4, No. 18 for details and elements of the equation)
$var_{two\text{-}g}$ (d_{g_t})	Variance estimate for two group design with repeated measures of the standardized mean difference proposed by Gibbons (See Table 4, No. 17 for details and elements of the equation)
Y^C	Control outcome
Y^E	Experimental outcome

Table 1. Standardized mean difference ES estimations (and their components) for a one-group repeated-measures design

No.	Source	Equation	Components
1.	Glass (1981)[3]	$d_{tra} = t_d \sqrt{\frac{1}{n} 2(1 - r_{Pre,Post})}$	$t_d = \frac{\overline{Y}_{Diff}}{\sqrt{\frac{2S_{Diff}^2\left(1 - r_{Pre,Post}\right)}{n}}} = \frac{\overline{Y}_{Pre} - \overline{Y}_{Post}}{S_{Diff}\sqrt{\frac{2\left(1 - r_{Pre,Post}\right)}{n}}}$ $S_{Diff} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i^{Diff} - \overline{Y}_{Diff}\right)^2}{n-1}} = \sqrt{S_{Pre}^2 + S_{Post}^2 - 2r_{Pre,Post}S_{Pre}S_{Post}}$ S_{Diff} = Standard deviation of the difference assuming unequal variances. n = number of observations $\overline{Y}_{Pre}$ = pretest mean of measure Y. $\overline{Y}_{Post}$ = posttest mean of measure Y. $r_{Pre,Post}$ = correlation between Y_{Pre} and Y_{Post}.
2.	Rosenthal (1991)[9]	$d_{tch} = t_d \sqrt{\frac{1}{n}}$	$t_d = \frac{\overline{Y}_{Diff}}{S_{Diff}}\sqrt{n}$
3.	Becker (1988)[19]	$d_b = c(n-1)\frac{\overline{Y}_{Post} - \overline{Y}_{Pre}}{S_{Pre}}$	$c(n-1) = 1 - \frac{3}{4(n-1)-1}$ S_{Pre} = standard deviation of the pretest
4.	Gibbons et al. (1993)[37]	$d_g = c(n-1)\frac{\overline{Y}_{Diff}}{S_{Diff}}$	
5.	Huedo-Medina & Johnson	$d_{hw} = c(n-1)\frac{\overline{Y}_{Post} - \overline{Y}_{Pre}}{S_{within-pool}}$	$S_{Within-pool} = \sqrt{\frac{(n-1)S_{Pre}^2 + (n-1)S_{Post}^2}{n-1}} = \sqrt{S_{Pre}^2 + S_{Post}^2}$

Table 2. Standardized mean difference (SMD) ES estimations (and their components) for two independent groups.

No.	Source	Equation	Components
6.	Hedges (1981)[38]	$d_{hb} = c(N-2)\dfrac{\overline{Y}^{E}_{Post} - \overline{Y}^{C}_{Post}}{S_{Pooled}}$	$c(N-2) = 1 - \dfrac{3}{4(N-2)-1}$ $S_{Pooled} = \sqrt{\dfrac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}}$ $N = n_E + n_C$ S_E = posttest standard deviation of the experimental group. S_C = posttest standard deviation of the control group. $\overline{Y}^{E}_{Post}$ = posttest mean of the experimental group $\overline{Y}^{C}_{Post}$ = posttest mean of the control group
7.	Becker (1988)[19]	$d_b = c(N-2)\left[\dfrac{\overline{Y}^{E}_{Post} - \overline{Y}^{E}_{Pre}}{S^{E}_{Pre}} - \dfrac{\overline{Y}^{C}_{Post} - \overline{Y}^{C}_{Pre}}{S^{C}_{Pre}}\right]$	
8.	Gibbons et al. (1993)[37]	$d_g = c(N-2)\left[\dfrac{\overline{Y}^{E}_{Diff}}{S^{E}_{Diff}} - \dfrac{\overline{Y}^{C}_{Diff}}{S^{C}_{Diff}}\right]$	
9.	Huedo-Medina & Johnson	$d_{hw} = c(N-2)\left[\dfrac{\overline{Y}^{E}_{Post} - \overline{Y}^{E}_{Pre}}{S^{E}_{within-pool}} - \dfrac{\overline{Y}^{C}_{Post} - \overline{Y}^{C}_{Pre}}{S^{C}_{within-pool}}\right]$	
10.	Shadish et al. (1999)[36]	$d_{s1} = \dfrac{\overline{Y}^{E}_{Post} - \overline{Y}^{C}_{Post}}{S_{Pooled}}$	$S_{Pooled} = \sqrt{\dfrac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}}$ $S_E = \dfrac{\lvert \overline{Y}^{E}_{Pre} - \overline{Y}^{E}_{Post} \rvert \sqrt{n_l}}{\lvert t^{E}_{d} \rvert \sqrt{2(1 - r^{E}_{Pre,Post})}}$

No.	Source	Equation	Components
			$S_C = \frac{\lvert \overline{Y}^C_{Pre} - \overline{Y}^C_{Post} \rvert \sqrt{n_l}}{\lvert t^C_d \rvert \sqrt{2(1 - r^C_{Pre,Post})}}$
11.	Shadish et al. (1999)[36]	$d_{s2} = \frac{\overline{Y}^E_{Post} - \overline{Y}^C_{Post}}{S_{ANOVA}}$	$S_{ANOVA} = \sqrt{\frac{MSE_b + (tp-1)MSE_w}{tp}}$ MSE_b = between-subjects mean square error MSE_w = within-subjects mean square error tp = number of measured time points
12.	Shadish et al. (1999)[36]	$d_{s3} = \frac{\overline{Y}^E_{Post} - \overline{Y}^C_{Post}}{S_{ANCOVA}}$	$S_{\text{ANCOVA}} = \sqrt{\frac{MSEa(N-h-1)}{(1-r^2_{w-class})(N-h)}}$ $r_{w-class} = \sqrt{\frac{F_{\text{cov}}}{F_{\text{cov}} + (N-h-1)}}$ MSE_a = adjusted mean square error from the covariance analysis

Table 3. Estimates of sampling variance for the SMD ES in the one-group design with repeated-measures

No.	Metric	Equation
13.	Raw-score metric (Becker, 1988)[19]	$\mathrm{var}_{one-g}(d_b)=\left(\frac{2\left(1-r_{pre,post}\right)}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1+\frac{n}{2\left(1-r_{Pre,Post}\right)}d_b^2\right)-\frac{d_b^2}{\left[c\left(100-1\right)\right]^2}$
14.	Change-score metric (Gibbons et al., 1993)[37]	$\mathrm{var}_{one-g}(d_g)=\left(\frac{1}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1+nd_g^2\right)-\frac{d_g^2}{\left[c\left(n-1\right)\right]^2}$

Note. "Raw-score" implies having the measures' variability in the observations and does not imply the unstandardized mean difference (UMD).

Table 4. Estimates of sampling variances for the SMD ES from two-group designs with repeated-measures

No.	Variance estimate	Equation
15.	Raw-score metric for a total ES (Hedges, 1981)[38]	$\mathrm{var}_{two-g}(d_{b_t})=\left(\frac{1}{\tilde{n}}\right)\left(\frac{N-2}{N-4}\right)\left(1+\tilde{n}d_{b_t}^2\right)-\frac{d_{b_t}^2}{\left[c\left(N-2\right)\right]^2},\ \tilde{n}=\frac{n_E*n_C}{n_E+n_C}$
16.	Raw-score metric for a function of two ESs (Becker, 1988)[19]	$\mathrm{var}_{two-g}(d_b)=\mathrm{var}_{one-g}(d_b^E)+\mathrm{var}_{one-g}(d_b^C)$
17.	Change-score metric for a total ES (Morris & DeShon, 2002)[22]	$\mathrm{var}_{two-g}(d_{g_t})=\left(\frac{1}{2\left(1-r_{Pre,Post}^E\right)\tilde{n}}\right)\left(\frac{N-2}{N-4}\right)\left(1+2\left(1-r_{Pre,Post}^E\right)\tilde{n}d_{g_t}^2\right)-\frac{d_{g_t}^2}{\left[c\left(N-2\right)\right]^2}$
18.	Change-score metric for a function of two ESs (Morris & DeShon, 2002)[22]	$\mathrm{var}_{two-g}(d_g)=\mathrm{var}_{one-g}(d_g^E)+\mathrm{var}_{one-g}^C(d_g^C)$

Note. "Raw-score" implies having the measures' variability in the observations and does not imply the unstandardized mean difference (UMD).

Table 5. Estimates of statistics related to the unstandardized mean difference (UMD) for designs with two independent groups

No.	Metric	Equation
19.	Unstandardized mean difference (UMD) for two independent groups	$UMD = \bar{Y}^{E}_{Post} - \bar{Y}^{C}_{Post}$
20.	Standard error of the UMD	$SE_{UMD} = S_{Pooled}\sqrt{\frac{1}{n_E} + \frac{1}{n_C}}$
21.	Inverse variance of the UMD	$\frac{1}{Var_{UMD}} = \frac{n_E n_C}{S_{Pooled}(n_E + n_C)}$

Methods

General Design

- To simulate each individual study in the simulation, we considered a within-subjects design with one or two groups, experimental and control.
- The data sets from a collection of *k* single studies were randomly generated using commands from the statistical software R, version 2.14.1. Two independent normal distributions were simulated for conditions in part 1; and for the conditions in part 2 that do not require repeated-measures design, R's rnorm command was used. Two bivariate normal distributions were generated specifically for part 2 when repeated-measures two-groups conditions were simulated using mvnorm R command. The variance-covariance matrix was manipulated with the identity matrix being a particular condition when groups of scores are not correlated and homogeneous variances equal to one are assumed between groups and time measures; the appropriate matrix was generated to create heterogeneous distributions for each group and time measure. The distributions were modified in some conditions, as we describe below.
- The two bivariate normal distributions, with a homogeneous variance-covariance matrix,

  $$Y^{E} \sim \left[N\left(\begin{pmatrix} \mu^{E}_{Pre} \\ \mu^{E}_{Post} \end{pmatrix}, \begin{pmatrix} \sigma^{2}_{Pre} & \sigma_{Pre,Post} \\ \sigma_{Pre,Post} & \sigma^{2}_{Post} \end{pmatrix} \right) \right], \; Y^{C} \sim \left[N\left(\begin{pmatrix} \mu^{C}_{Pre} \\ \mu^{C}_{Post} \end{pmatrix}, \begin{pmatrix} \sigma^{2}_{Pre} & \sigma_{Pre,Post} \\ \sigma_{Pre,Post} & \sigma^{2}_{Post} \end{pmatrix} \right) \right],$$

  were generated, representing the experimental and control groups, respectively. The parameters for these distributions in the standardized units are $\mu^{E}_{Pre} = \mu^{C}_{Post} = \mu^{C}_{Pre} = 0$, with $\sigma_{Pre,Post}$, σ^{2}_{Pre}, σ^{2}_{Post}, and μ^{E}_{Post} being manipulated factors in the simulation. These values were permitted to remain in their unstandardized units to create a comparison for statistical inferences between raw and standardized conditions.
- The necessary basic statistics (i.e., means, standard deviations, correlations) were estimated from the sampling data for each method using basic R commands. Thus, the estimates of ES and the ES sampling variance were calculated programming in R all the effect size equations from Tables 1, 2, and 5, as relevant.
- The calculations for the estimations, and their sampling variances were repeated for each simulated study (and comparing the equations in Tables 1-5; note that the between-groups sampling variance is the same as the raw-score metric, that is, Table 4, no. 15).
- In order to evaluate the robustness of the estimates under different conditions the percentage bias of the estimate was calculated:

  $$Bias\left(\hat{\delta}\right) = \frac{\sum_{j=1}^{Rns}\left(\frac{\hat{\delta}_j - \delta}{\delta}\right)}{Rns},$$

  where $\hat{\delta}_j$ is the sample estimate of population parameter δ for the j_{th} replication and *Rns* is the number of replications.
- The efficiency of the estimates was obtained as the variability of the estimate across replications,

$$VAR(\hat{\delta}) = \frac{\sum_{j=1}^{Rns}\left(\hat{\delta}_j - \delta\right)^2}{Rns}.$$

- Finally, the particular formulas for estimating the sampling variances of each index of specific aim 2 were computed in each replication and their values averaged over the 10,000 replications of the same condition. The average of the empirical variability of each index was compared to the average of the variance obtained from each sampling variance estimate to obtain the adjustment to the theoretical variance.

Conditions Manipulated

Table 6 summarizes all manipulated features in the simulation. These included:

1. The number of studies in a meta-analysis, k = 10, 20, 50, and 100.
2. The mean sample size (N) in the literature. The mean sample size for each generated meta-analysis replicated the 10th (N = 30), 40th (N = 50), and 80th (N = 80) percentiles of the sample sizes from HIV prevention trials in the Syntheses of HIV and AIDS Research Project (SHARP) meta-analytic database at the University of Connecticut, which summarizes over 700 trials. Three vectors of sample sizes were generated as [12, 16, 18, 20, 84], [32, 36, 38, 40, 104], and [62, 66, 68, 70, 134], one for each selected averaging 30, 50, and 80, respectively. Each vector was replicated either 2, 4, 10, or 50 times for meta-analyses of k = 10, 20, 50, and 100.
3. The within-study variances for experimental and control groups and at pretest and posttest measures were varied using ratios for experimental and control groups, respectively, of 1:1, 2:1, and 4:1.[26,27] The variance of the experimental group was increased in comparison to that of the control group because increases in variability are more plausible when there is experimental manipulation (e.g., a psychological treatment) and doing so permitted clearer inferences about results.[3]
4. The mean of the posttest for the experimental group, following the parametric values for the standardized mean difference,[28] $\delta = \mu^{E}_{post}$ = 0.25, 0.5, 0.8, The means and standard deviations of the scores for the experimental and control participants in each pseudo-study were generated assuming a variety of different distributions: both normal distributions and non-normal distributions:
    a. For the normal distributions, values for means and standard deviations were kept as following the parametric normal distributions described above.
    b. To generate non-normal distributions, the normality pattern was manipulated to obtain skewed distributions through use of the Fleishman[29] algorithm, with the following values of skewness/kurtosis: 0.5/0, 0.75/0, and 1.75/3.75..
5. The between-studies variance, τ^2, with values 0, 0.04, 0.08, 0.16, 0.32. When $\tau^2 = 0$, the statistical model reduces to a fixed-effects model, because there is no between-studies variance. The selected values of τ^2 are similar to those used in other prominent simulation studies in this literature.[30-35]

Condition manipulated specifically for Specific Aim 2

6. The correlations between the two conditions were manipulated through the variance-covariance matrix, where homogeneous variances equal to 1 can be assumed. The manipulated correlation was equal to the covariance between the two measures. The values were $\rho_{pre\text{-}post}$ = 0, 0.25, 0.50, and 0.75.

Table 6. Characteristics of Simulated Datasets

Condition	Levels
Mean population effect size	$\delta = \mu^{E}_{post}$ = 0.25 / 0.50 / 0.80
Relative within-study standard deviation for control and experimental groups	$\sigma_E : \sigma_C$ = 1:1, 2:1, and 4:1
Sample sizes vectors	$n_C = n_E$ = [12, 16, 18, 20, 84], [32, 36, 38, 40, 104], and [62, 66, 68, 70, 134]
Number of studies	k = 10 / 25 / 50 / 100
Skewness/Kurtosis	0/0, 0.5/0, 0.75/0, and 1.75/3.75
Between-study variance	τ^2 = 0 / 0.04 / 0.08 / 0.16 / 0.32
Correlation between pretest and posttest measures[a]	$\rho_{pre\text{-}post}$ = 0 / 0.25 / 0.50 / 0.75

[a]This portion of the design was only for Specific Aim 2

Results

Part One: Unstandardized Versus Standardized Effect Sizes

Overview

Monte Carlo simulations showed that leaving the effect size (ES) index in the original metric (UMD) presents little bias or loss of efficiency when distributions were normal, when there is no heterogeneity in effect sizes, and when the variances of the experimental and control group means are equal; yet, to the extent that these conditions deviate, standardizing (SMD) is better. The standardized metric presents the least bias under all conditions and is more efficient than the raw metric. Both metrics suffer under high skewness and kurtosis, although the SMD less so.

Detailed Analysis

Bias

Figures 1 through 5 show the primary results comparing the UMD against the SMD in terms of bias. As these Figures show, the SMD had less bias than did the UMD under all plotted circumstances. The two metrics approach the same level of bias only when skewness and kurtosis is minimal (Figure 1), but even here the SMD showed a slight advantage. More dramatic differences between the two appeared as skewness and kurtosis increase, under fixed-effects (τ^2=0; see Figure 2), as the parametric effect size increases (Figure 3), as the mean sample size

increases (Figure 4), and as heteroskedasticity increases (Figure 5). (No figure for number of studies appears because it did not change the trends reported here.) Across these conditions, the UMD was more likely than the SMD to over-estimate the parametric effect size.

Figure 1. Bias of SMD and UMD as a function of asymmetry and the between-studies variance of the distribution.

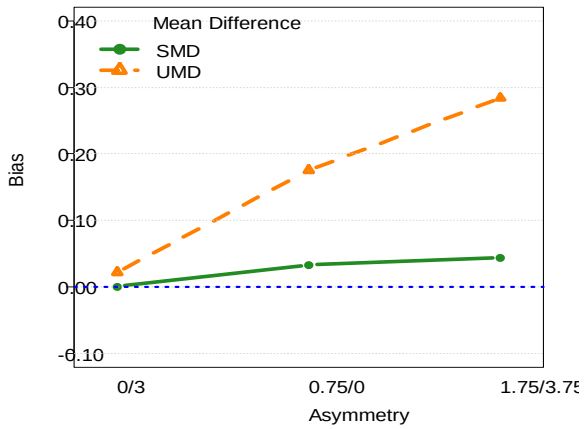


Figure 2. Bias of SMD and UMD as a function of the between-studies variance of the distribution.

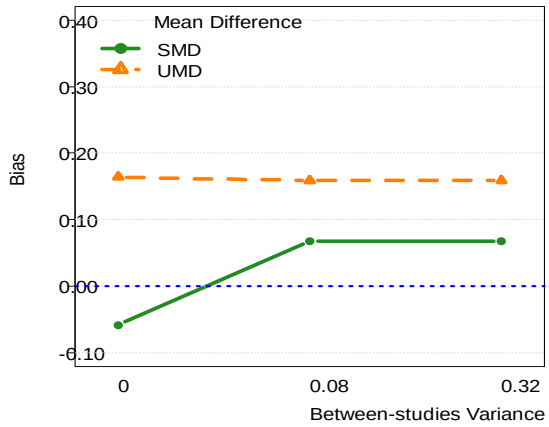


Figure 3. Bias of SMD and UMD as a function of the parametric effect size.

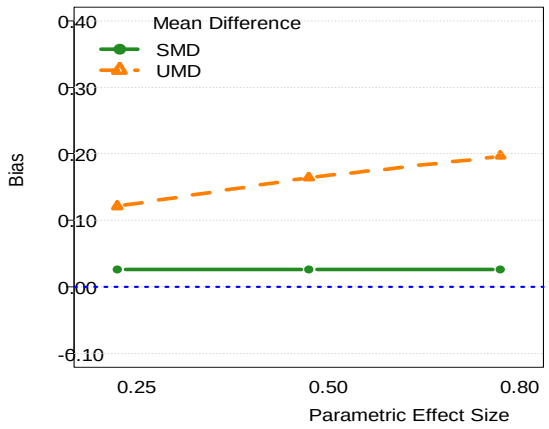


Figure 4. Bias of SMD and UMD as a function of mean sample size.

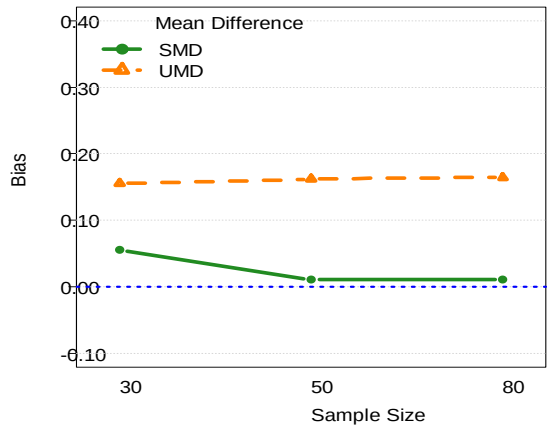
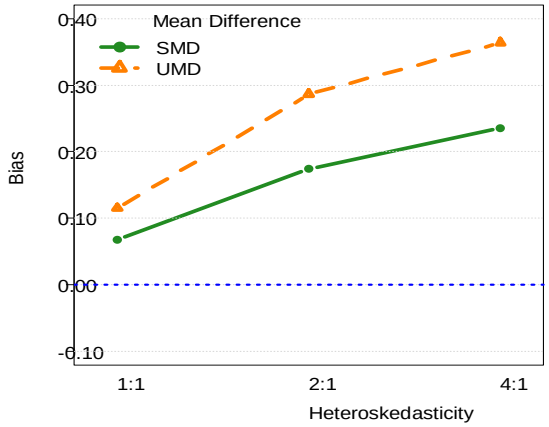


Figure 5. Bias of SMD and UMD as a function of heteroskedasticity.



Efficiency

Figures 6 through 10 show the primary results comparing the UMD against the SMD in terms of efficiency. The SMD presented more efficient estimations than the UMD under the same circumstances as it exhibited less bias. Specifically, the two metrics approach the same efficiency only when skewness and kurtosis is minimal (Figure 6), but even here the SMD showed a slight advantage. More dramatic differences between the two appeared as skewness and kurtosis increase, under increasing heterogeneity ($\tau^2>0$; see Figure 7), across all levels of the parametric effect size (Figure 8), across all levels of mean sample size (Figure 9), and as heteroskedasticity increases (Figure 10). (No figure for number of studies appears because it did not change the trends reported here.) Across these conditions, the UMD was a less efficient estimator than the SMD. The efficiency of SMD improves when the between-studies variance increases ($\tau^2>0$) or the sample size increases; it is not affected by the rest of the factors.

Figure 6. Efficiency of SMD and UMD as a function of asymmetry.

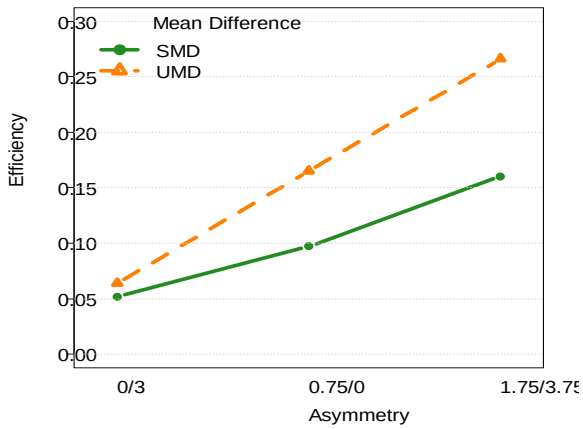


Figure 7. Efficiency of SMD and UMD as a function between-studies variance.

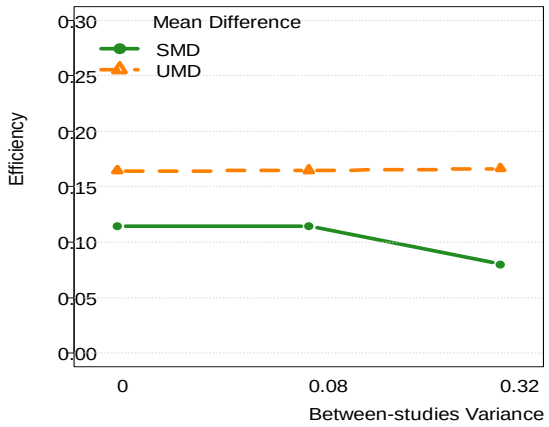


Figure 8. Efficiency of SMD and UMD as a function of the parametric effect size.

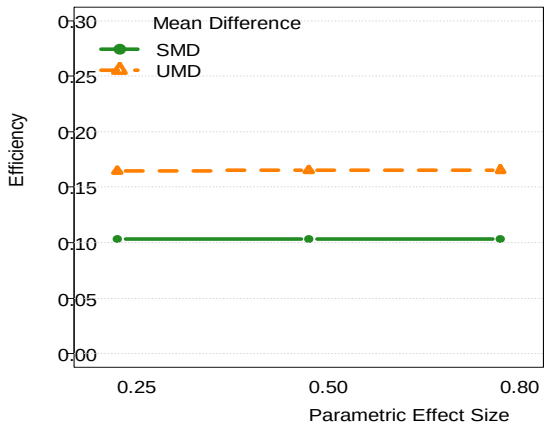


Figure 9. Efficiency of SMD and UMD as a function of mean sample size.

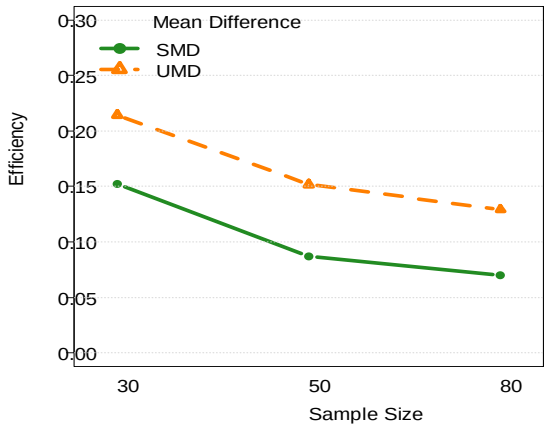
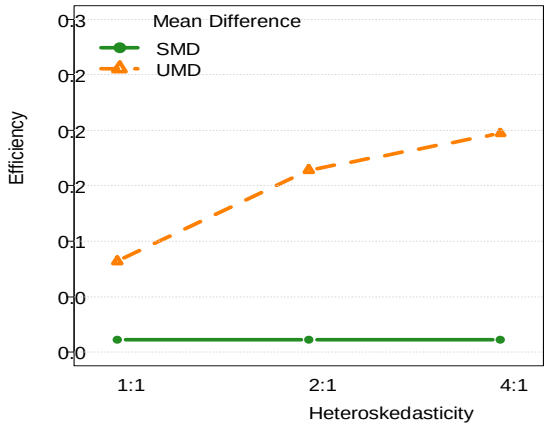


Figure 10. Efficiency of SMD and UMD as a function of heteroskedasticity.



Part Two: Using the SMD with Repeated Versus Independent Measures

Overview

- The raw-score metric is more convenient if we are interested in analyzing just the effect of an intervention, excluding effects of the maturation (i.e., when a process within an individual can vary due to the passage of time and/or learning independently of specific external events, for example, growing older, wiser, and more experienced). Therefore, when the raw-score metric is, used the consistency of the scores has not been considered and post hoc analyses including this factor as a moderator should not be conducted. However, the change-score metric allows us to know not only the change due to the intervention but also if there is some consistency due to a maturation effect. Examining maturation effects is very appropriate in some educational and even clinical interventions and if is included in the ES, then the consistency effect in the final efficacy results can be evaluated. (NB. "Raw-score metric" should not be confused with the unstandardized mean difference. The former gauges the standardized differences between two groups taking into account their variability but not the correlation between two measures. Part Two considers only the standardized mean difference, SMD.)
- There are important differences among the ES estimations in raw-score and in change-score metric. However, all of them can be comparable and combined if it is possible to assume a moderate correlation (r=0.50) between pretest and posttest measures.
- The new ES estimation proposed in this work, d_{hw}, which divides the difference of the means, pretest and posttest, by a function of their variances is very efficient but it should be used only when the researcher can assume independence between the two measures. Therefore, this solution is more convenient to use for repeated-measures two-group designs focusing on intervention effects.
- There are some slight differences between the empirical variances of raw- and change-score metrics when the correlation is larger or smaller than 0.5.
- The weighted analyses using the different metrics are comparable because the differences due to the correlation factor are corrected using the appropriate weights in each case. Thus, the correlation factor should be included for the sampling variance but not for the ES estimation in raw-score metric and should not be included for the sampling variance but for the ES estimation in change-score metric. Therefore, those ES values in raw-score metric with a greater sampling variance will have less weight in change-score metric but the ES values will be larger than the one in raw-score metric if the correlation is larger than 0.5. In the same way, the weights will be larger in raw-score metric ESs when the correlation is larger than 0.5 but the ES values will be smaller than those in change-score metric.
- Therefore in order to be able to integrate both metrics, the sampling variance of the ES estimations in raw-score metric must include the factor with the correlation, but not the sampling variance estimation for change-score metric. However, if we can assume that the correlation is 0.5, both equations to estimate the sampling variance as it was for the ES estimations are comparable.

- For the repeated-measures two-groups design, the sampling variance can be estimated using two equations depending on the metric of the ES. The main difference between them is the inclusion or not of the correlation factor. The one for change-score metric includes the correlation factor but not the one in raw-score metric. The performance of those two equations presents the same trends as those for the one-group within-subjects design. When $\rho < 0.5$ the one for change-score metric is better adjusted than the one for raw-score metric, which is better, regardless of the ES index used to calculate it, when $\rho > 0.5$. They are comparable if it is possible to assume that the correlation is equal to 0.5.

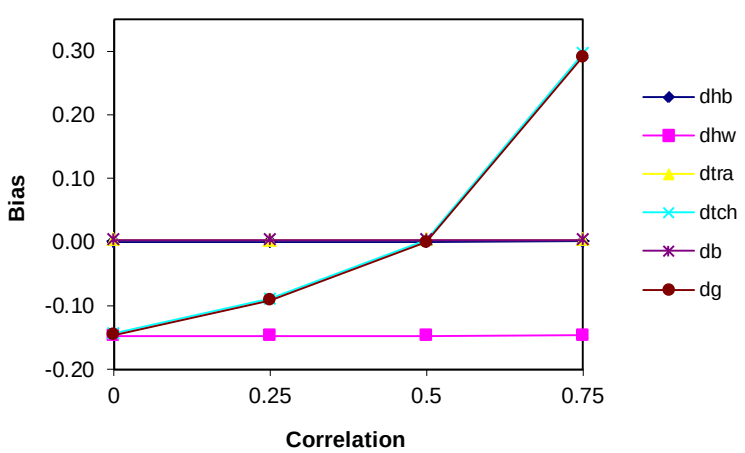
Detailed Analysis

One-group within-subjects design

Bias

Figure 11 shows the bias of the different indexes calculated as the difference between the mean of each estimation and the parametric ES in raw metric, δ_{raw}, over the 10,000 replications and assuming normality of the distribution. Thus, positive values reflect an overestimation of this particular parameter, whereas negative values imply an underestimation of it. The trends were very similar as the sample size increases, so an average of the sample size conditions is used to present each ES index across each manipulated correlation value.

Figure 11. Bias as a function of the correlation parameter of the ES indexes.



Note: Hedges's standardized mean difference using the between-studies degrees of freedom=d_{hb}, Hedges's standardized mean difference using the within-study degrees of freedom=d_{hw}, standardized mean difference in raw-score metric from t_d=d_{tra}, the standardized mean difference in change-score metric from t_d=d_{tch}, Becker's standardized mean difference=d_b; Gibbons's standardized mean difference=d_g.

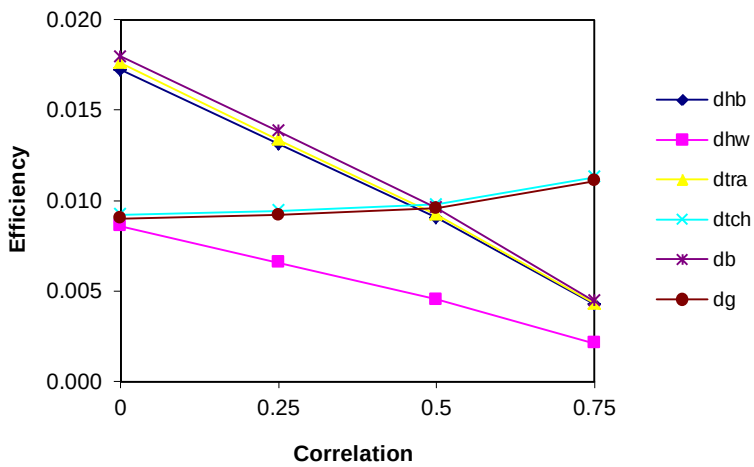
The ES indexes in raw-score metric do not present bias, d_{hb}, d_{tra}, and d_b, except for d_{hw}, which underestimates the parametric ES in raw-score metric and when the ES is estimated assuming the two groups of scores have a null correlation, $\rho = 0$, which is similar in all the conditions (see Appendix B for more detail). However, the ES indexes in change-score metric, d_g and d_{tch}, which incorporate the effect of the correlation, underestimate the parametric ES in raw-score metric when $\rho < 0.5$.

It is possible to observe the interaction of all the ES estimations, except d_{hw}, when $\rho = .5$. They are estimating the same parameter for this particular condition, so in this case d_{hb}, d_{tra}, d_{tch}, d_b, and d_g are unbiased.

Efficiency

In terms of the theoretical variance, the ES with the least variability and therefore most efficient is d_{hw} across all simulated conditions, then d_g and d_{tch}, both presenting the same trend, as Figure 12 shows.

Figure 12. The efficiency as a function of the correlation parameter of the ES indexes



Note: Hedges's standardized mean difference using the between-studies degrees of freedom=d_{hb}, Hedges's standardized mean difference using the within-study degrees of freedom=d_{hw}, standardized mean difference in raw-score metric from t_d=d_{tra}, the standardized mean difference in change-score metric from t_d=d_{tch}, Becker's standardized mean difference=d_b; Gibbons's standardized mean difference=d_g.

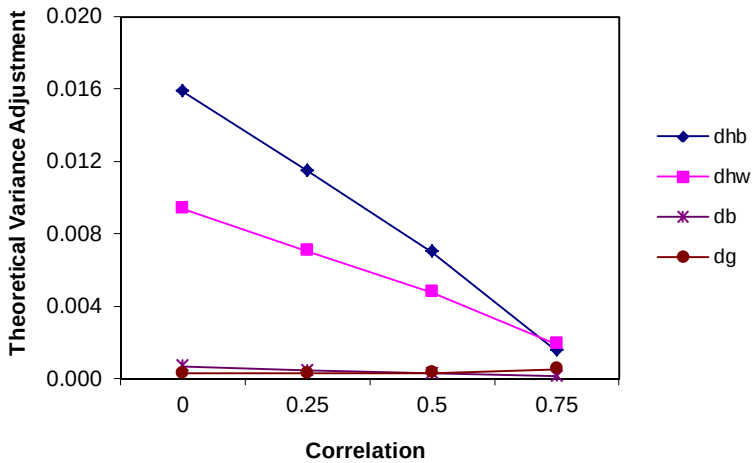
The variability of d_g and d_{tch} are almost unaffected by increasing the assumed correlation between pre- and post-test, although it increases slightly as the correlation increases. The variability of d_{hw} begins very similar to d_g and d_{tch}, but the efficiency of the first one improves as the correlation increases. Finally, the worst adjustment is for d_b, followed by d_{tra} and d_{hb}, however, their variability decreases as the correlation increases, obtaining better efficiency than d_g when $\rho > 0.5$. Although efficiency is similar for all alternatives when the correlation is 0.5, d_g and d_{tch} are more variable than any other when the correlation is larger than 0.5. As statistical theory predicts, the performance of all ESs, in terms of efficiency, improves as the sample size increases. Finally, in general, the correlation affects the empirical variance of the estimators even for those in raw-score metric. As the correlation increases, there is less variability in the indexes.

The theoretical variance adjustment

Related to the difference between the empirical and the theoretical variance, in general, Figure 13 shows how the correlation parameter of the ES indexes relates to adjustments to the theoretical variance. The least biased estimations of the variance are d_g and d_b. The difference between the theoretical and the empirical variance decreases slightly for d_g as the correlation increases if the latter is smaller than 0.5, however, an opposite pattern is observed if the

correlation is larger than 0.5. However, the theoretical variance adjustment of d_b always decreases as the correlation increases. The worst adjustment is for d_{hb} and then for d_{hw}; both present a difference that decreases as the correlation increases. Finally, d_b and d_g have well-adjusted estimations of the sampling variance when $n > 100$.

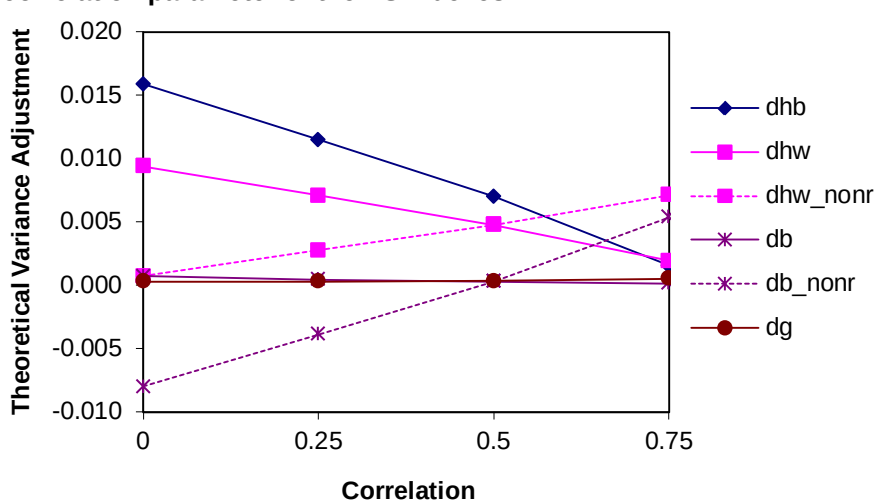
Figure 13. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes.



Note: Hedges's standardized mean difference using the between-studies degrees of freedom=d_{hb}, Hedges' standardized mean difference using the within-study degrees of freedom=d_{hw}, Becker's standardized mean difference=d_b, and Gibbons's standardized mean difference=d_g.

If the theoretical variance is calculated for d_{hw} and d_b without including the correlation factor, $2(1 - \rho)$, but using the $df = n - 1$ and $\tilde{n} = n$, d_{hw_nonr} and d_{b_nonr}, they present opposite patterns to their correspond versions including the correlation factor, as Figure 14 illustrates.

Figure 14. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes.



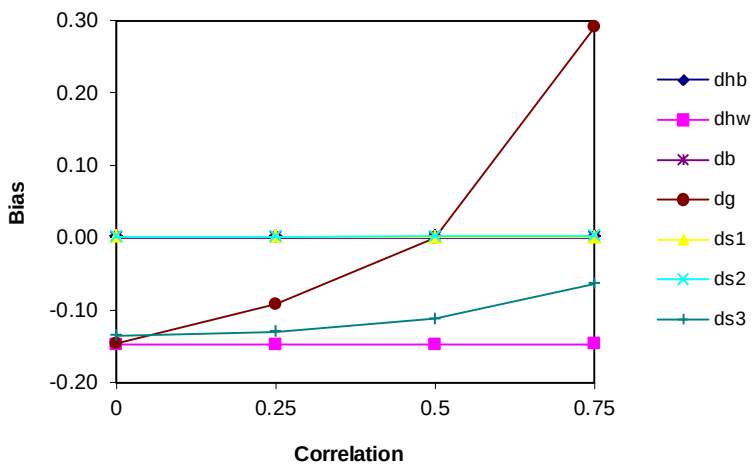
Note: Hedges's standardized mean difference using the between-studies degrees of freedom=d_{hb}; Hedges's standardized mean difference using the within-study degrees of freedom=d_{hw}; Hedges' standardized mean difference using the within-study degrees of freedom without including the correlation factor 2(1 – *r*) in its equation=d_{hw_nonr}; Becker's standardized mean difference=d_b; Becker's standardized mean difference without including the correlation factor 2(1 – r) in its equation=d_{b_nonr}; and Gibbons' standardized mean difference=d_g.

Two-Group Repeated Measures Design

Bias

In Figure 15, the same trends as in Figure 11 are presented for the correspondent ES indexes of two-group repeated-measures design, d_{hb}, d_b, d_{hw} and d_g. The two first are estimating the parametric ES in raw-score metric, so they are not presenting bias, in comparison to the constant bias of d_{hw} across conditions because it is not affected by the correlation but it is not estimating their parametric change-score ES (the mean of the bias through all conditions is bias(d_{hb}) = 0.0016; bias(d_b) = 0.000; bias(d_{hw}) = -0.1473). The d_g ES presents the same result mentioned under one-group repeated-measures design, which is underestimating and overestimating the parametric ES depending on having $\rho < 0.5$ or $\rho > 0.5$, respectively.

Figure 15. The bias as a function of the correlation parameter of the ES indexes.



Note: Hedges's standardized mean difference using the between-studies degrees of freedom=d_{hb}; Hedges's standardized mean difference using the within-study degrees of freedom=d_{hw}; the Becker's standardized mean difference=d_b; the Gibbons' standardized mean difference, d_g, the standardized mean difference using the Shadish's pool standard deviation= d_{s1}; the standardized mean difference from the ANOVA data, d_{s2}; and the standardized mean difference from ANCOVA data=d_{s3}.

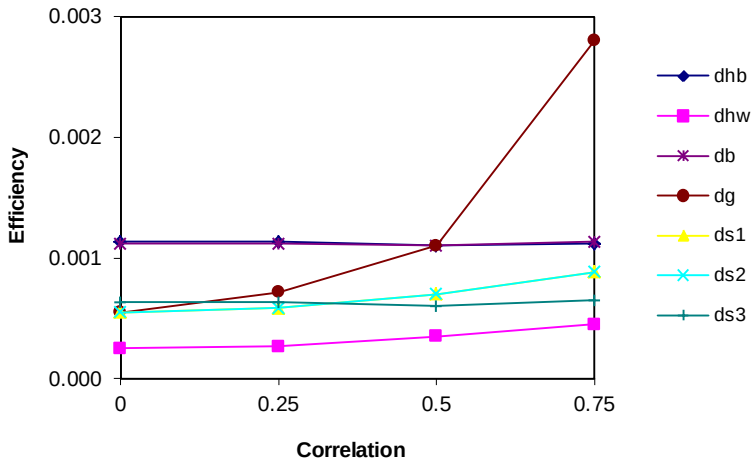
The three new ES indexes presented for two-groups design are those calculated from ANOVAs, d_{s1} and d_{s2}, and one from ANCOVA, d_{s3}. The performance of the two first is the same as those estimating a parametric ES in raw-score metric, so they do not present bias. However, the third one underestimates the parameter but improves as the correlation increases.

Efficiency

In general, in terms of efficiency the best estimator is d_{hw} and then d_{s1}, d_{s2}, d_{s3}, and d_g are very similar if correlation is lower than 0.5. However, when $\rho > 0.5$ the efficiency of d_g increases drastically, as it was shown in one-group design. The variability for d_{hb} and d_b are very similar,

being slightly larger for d_b when $\rho = 0$, they interact when $\rho = 0.5$ being from that point the variability of d_{hb} slightly larger than d_b (see Figure 16).

Figure 16. Efficiency as a function of the correlation parameter of the ES indexes.

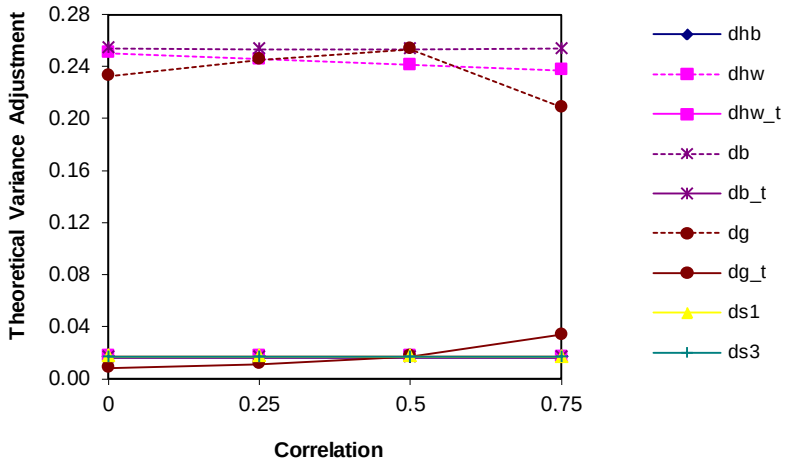


Note: Hedges's standardized mean difference using the between-studies degrees of freedom=d_{hb}; Hedges's standardized mean difference using the within-study degrees of freedom=d_{hw}; Becker's standardized mean difference=d_b; Gibbons's standardized mean difference=d_g; standardized mean difference using the Shadish pooled standard deviation=d_{s1}; standardized mean difference from the ANOVA data=d_{s2}; and standardized mean difference from ANCOVA data=d_{s3}.

The theoretical variance adjustment

Figures 17 and 18 present the differences between the sampling variance estimates and the actual sampling variance of the effect size index. In Figure 17, the estimates using a continuous line are calculated as the variance of an ES, those with a dashed line are obtained as a variance compound of two variances, one for the ES of each independent group.

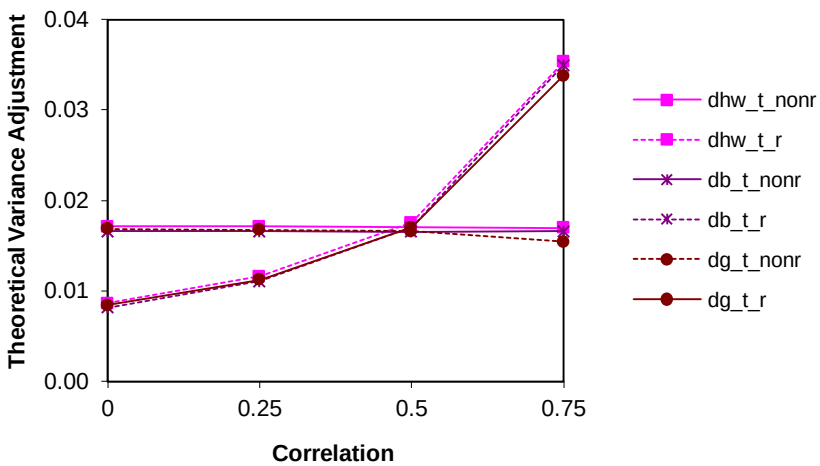
Figure 17. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes.



Note: Hedges's standardized mean difference using the between-studies degrees of freedom, d_{hb}, Hedges' standardized mean difference using the within-study degrees of freedom, d_{hw}, Becker's standardized mean difference, d_b, Gibbons' standardized

mean difference, d_g, standardized mean difference using the Shadish pooled standard deviation, d_{s1}, standardized mean difference from the ANOVA data, d_{s2}, and the standardized mean difference from ANCOVA data, d_{s3}; note that the variance calculated for a total ES as the difference between the ES for each group, experimental and control, is denoted with *_t* suffixes.

Figure 18. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes.



Note: Hedges's standardized mean difference using the within-study degrees of freedom, d_{hw}; Becker's standardized mean difference, d_b; Gibbons's standardized mean difference=d_g; note that the inclusion or not of the correlation factor, 2(1-ρ), is denoted as *_r* and *_nonr*, respectively, in the suffix of these terms

The indexes that are adjusted most through all the conditions without being affected by the magnitude of the correlation are d_{hb} and $d_{b_total,}$ and d_{hw_t} . The adjustments for the theoretical variance of d_{g_t}, d_{s1}, d_{s2}, and d_{s3} are affected by the magnitude of the correlation between pre- and post-tests (d_{s2} is not considered in Figure 17 because it has exactly the same performance as d_{s1}). The theoretical variance is closer to the empirical variance for d_{g_t}, d_{s1}, and d_{s2} as the correlation increases, but it is slightly further for d_{s3} when $\rho > 0.5$. The considerably worse performance is achieved by those estimations of the sampling variance that are a composite of two variances, one for the experimental and one for the control group.

In Figure 18, the sampling variance estimations for a total ES in different metrics are compared using or not using the correlation factor, $2(1 - \rho)$. The performance has a good adjustment almost without being affected by the increase of the correlation for all the estimations and they are very similar when they do not include the correlation factor. The only one that improves slightly as the correlation decreases is the estimation of the sampling variance using d_g. However, when the correlation factor is included in the equation all of them have similar trends, they get better performance if $\rho < 0.5$ and worse performance when $\rho > 0.5$. Finally, if $\rho = 0.5$ the inclusion or not of the correlation factor has little effect across the different equations.

Discussion

This report analytically evaluates two controversial and related topics in meta-analytic methods using Monte Carlo simulation techniques. The first is to determine what effect size metric should be used when trials assess an outcome on the same measures. The second is to determine the best estimates of the standardized effect size and its variance estimates when the comparisons are derived from a repeated-measures or between-groups design.

Choice of Metric

Although several statistical methods exist to estimate comparisons of groups at one or more points (Tables 1 to 4), none provide unbiased estimations and, before the current report, the circumstances under which they produce the most optimal statistical inferences has been unknown. The current results demonstrate that the standardized version of effect size outperforms the unstandardized version under a broad set of conditions (Table 6) in terms both of bias (Figures 1-5) and of efficiency (Figures 6-10). Thus, standardized effect sizes can be recommended under most conditions. The standardized mean difference performs much better when within-study variability is large, when parametric assumptions are poorly met, and when study sample sizes are small. Table 6 summarizes the recommendations that the current simulation supports in terms of effect size metric and the best instantiations of the standardized mean difference for particular types of designs and inferential circumstances.

The fact that the current results support the use of the standardized mean difference even when it is possible to use the unstandardized version might on the surface imply that that clinical interpretations will grow more difficult even while statistical inferences grow clearer and cleaner. Of course, most stakeholders can more easily interpret a 10 mmHg greater drop in blood pressure or a $100 reduction in the cost of care than the equivalent result on a standardized effect size metric. There are at least two solutions to this problem. The first solution is quantitative and entails converting final results from in the standardized mean difference metric to their equivalent unstandardized mean differences. One simply multiplies the standardized mean difference by the standard deviation. Of course, standard deviations can and do vary widely between studies, which implies that is valuable to meta-analyze the relevant standard deviations in order to determine which value or values are best used in such conversions. Many factors might affect which standard deviation is presumed to describe a particular inferential situation. Investigators may have selected participants within a narrow range on the dependent measure, which artificially restricts the standard deviation. Presumably such standard deviations are of little use in setting a standard. Scaling issues are also a consideration: Other factors equal, standard deviations will grow smaller as values near the low or high extremes of a particular measure (e.g., rating scales); standard deviations grow larger across levels of a measure that has infinity at one end (e.g., mmHg in blood pressure studies).[6] Understanding when the standard deviation is larger or smaller thus facilitates making accurate clinical inferences.

Table 7. Recommendations for Effect Size Metric and Relevant Formulas

Design Type	Inferential Circumstances	Best Effect Size Statistic	Source	Equation
Two-group comparison without repeated measures	For nearly all inferential circumstances involving a comparison of two groups on one measure.	SMD d_{hb},Table 2, No. 6	Hedges (1981)[38]	$d_{hb} = c(N-2)\frac{\overline{Y}^{E}_{Post} - \overline{Y}^{C}_{Post}}{S_{Pooled}}$
One-group repeated measures	When attempting to describe the magnitude of change without controlling for the correlation between the repeated measures	SMD d_{tra}, Table 1, No. 1	Glass (1981)[3]	$d_{tra} = t_d\sqrt{\frac{1}{n}2(1-r_{Pre,Post})}$
		SMD d_b, Table 1, No. 3	Becker (1988)[19]	$d_b = c(n-1)\frac{\overline{Y}_{Post} - \overline{Y}_{Pre}}{S_{Pre}}$
One-group repeated measures	When attempting to examine the magnitude of change controlling for the correlation between the repeated measures	SMD d_{tch}, Table 1, No. 2	Rosenthal (1991)[9]	$d_{tch} = t_d\sqrt{\frac{1}{n}}$
		SMD d_g, Table 1, No. 4	Gibbons et al. (1993)[37]	$d_g = c(n-1)\frac{\overline{Y}_{Diff}}{S_{Diff}}$
Two-group comparisons with repeated measures	When attempting to examine the magnitude of change without controlling for the correlation between the repeated measures	SMD d_b, Table 2, No. 7	Becker (1988)[19]	$d_b = c(N-2)\left[\frac{\overline{Y}^{E}_{Post} - \overline{Y}^{E}_{Pre}}{S^{E}_{Pre}} - \frac{\overline{Y}^{C}_{Post} - \overline{Y}^{C}_{Pre}}{S^{C}_{Pre}}\right]$
		SMD d_{hw}, Table 2, No. 9	Huedo-Medina & Johnson	$d_{hw} = c(N-2)\left[\frac{\overline{Y}^{E}_{Post} - \overline{Y}^{E}_{Pre}}{S^{E}_{within-pool}} - \frac{\overline{Y}^{C}_{Post} - \overline{Y}^{C}_{Pre}}{S^{C}_{within-pool}}\right]$
Two-group comparisons with repeated measures	When attempting to examine the magnitude of change controlling for the correlation between the repeated measures	SMD d_g, Table 2, No. 8	Gibbons et al. (1993)[37]	$d_g = c(N-2)\left[\frac{\overline{Y}^{E}_{Diff}}{S^{E}_{Diff}} - \frac{\overline{Y}^{C}_{Diff}}{S^{C}_{Diff}}\right]$
		SMDs d_{s2} or d_{s3}, Table 2, Nos. 11 and 12	Shadish et al. (1999)[36]	$d_{s2} = \frac{\overline{Y}^{E}_{Post} - \overline{Y}^{C}_{Post}}{S_{ANOVA}}$ $d_{s3} = \frac{\overline{Y}^{E}_{Post} - \overline{Y}^{C}_{Post}}{S_{ANCOVA}}$

The second solution for clinical interpretation hinges on effect size standards. Specifically, Cohen[28,39] tentatively proposed some guidelines for judging effect magnitude, suggesting "that medium represents an effect of a size likely to be visible to the naked eye of a careful observer" (Cohen,[24] p. 156). Thus, if a standardized mean difference exceeds 0.50, then it is likely to be readily noticeable to the careful practitioner.[40] If it is smaller, it is unlikely to be noticeable without the aid of statistics. It is worth noting that these clinical interpretation suggestions also apply to meta-analyses in which individual studies take observations on different measures, when the only conventional recourse is to use a standardized effect size.

Optimal Estimations of the Standardized Mean Difference Effect Size (and Its Sampling Variance)

Tables 1 to 4 show current methods to obtain an effect size and a sampling variance estimate for repeated-measures and two-groups designs. These solutions either include the correlation between pre- and post-test [3,21,36] or exclude it.[9,19,36,37] Despite the disagreement about use of the correlation in calculating the ES, all except Gibbons et al.'s[37] solutions use the correlation in estimating the variance of ES for subsequent weighted analyses. Finally, these solutions rarely if ever distinguish between change- and raw-score metrics; the latter always assumes a .5 correlation between measures in estimating the ES and its variance. The effect size in change-score metric can be defined as the mean change due to treatment compared to the variability of change scores and the effect size in raw-score metric as the mean difference between conditions compared to a pooled variability of scores within each condition or to the variance of the original scores without having any intervention.

The second takes into account only the change, without considering the variability of this change, and the first considers the change and the consistency of it. If the variability of this change is high, the ES in change-score formulation will be smaller than it will be in the raw-score formulation, that is considering just the between groups variability implying that the correlation between the two conditions is 0.5. Thus, the raw-score ES can be misleading. However, if the variability of the change is small, the ES estimation will be higher than if just the ES in raw-metric is considered because of the consistency. Consistency implies that for all the subjects, a similar change has been produced. Therefore, those metrics will report different definitions of the ES because of the different standard deviations that they use.

There are different estimates of the sampling variance depending on the design (Tables 3 and 4); all present a good adjustment to the theoretical variance under most circumstances. Yet, for two-groups designs with repeated measures, there is an advantage to use the equations with the total effect size as a component (i.e., Table 4, equations 15 and 17). These performed superior to versions that used separate variance estimates for the two compared groups to create the total sampling variance (i.e., Table 4, equations 16 and 18).

We can conclude based on our results that selections of a formula for repeated measures can have large effects on statistical inferences. The parametric repeated-measures ES is defined as the difference between the means of the post- and pre-test divided by a standard deviation. The particular standard deviation chosen in calculating the ES index will also create some differences. Those differences can be corrected using the appropriate weights in each case, using the sampling variance estimate for change- or raw-score metric, then effect sizes from different designs can be integrated. It is worth mentioning that solutions for repeated measures effect sizes were most optimal when the correlation between repeated observations was 0.50; to the extent

that actual observed correlations differ from this value, statistical inferences are likely to be sub-optimal, especially with some of the competing equations (Tables 1 and 2).

Limitations and Future Directions

The present study examined the performance of numerous estimators of effect size across widely diverse circumstances but it cannot evaluate all possible circumstances. Although the methods were intended to describe the conditions that most often appear in meta-analyses of health-related research, it is possible that important conditions have been omitted from the current simulations. For example, trials sometimes have far larger samples than the current simulations examined. Yet, because sample size had little role in results, this concern would seem to be abated. Moreover, in examining circumstances with heterogeneity and with unequal variances, the current findings would seem highly germane to many meta-analyses related to health.

The current investigation also leaves some questions without complete answers. Future investigations could examine alternative solutions beyond those in Table 5 for gauging the magnitude of effect sizes in the original metric. For example, as implied in the preceding sub-section, it may be fruitful to model the standard deviations in trials. Once the population values are estimated they could be used in place of the observed standard deviations in individual studies to weight results. This solution might correct many of the deficiencies the current study identified. (Or, the population standard deviations could replace the observed standard deviations in calculating the standardized mean difference.) Another solution could be taking previous transformations of the unstandardized metric and evaluating which ones are the most unbiased and efficient depending on different simulated conditions. Similarly, in comparing the unstandardized effect size to the standardized one, the current work examined only one version (see Table 5). One popular version that was not examined in the current analysis is the unstandardized mean gain score.[16]

The unstandardized difference's poor performance in the current analysis leaves little faith that it will fare any better in the gain score arena, but only by doing the requisite work can this possibility be confirmed. Similarly, the current finding that the standardized mean difference performs better than the unstandardized one under unequal variances implies but does not directly show that differing measure variances across studies will make the unstandardized mean difference perform more poorly. Moreover, the current results showed that the standardized mean difference performs better under heterogeneity than its unstandardized counterpart; the implication is that moderator testing (viz. sub-group analysis or meta-regression) will also exhibit less bias and greater efficiency when the effect size is standardized rather than unstandardized. This possibility should be evaluated in a future simulation. Another important possibility to evaluate in a future study are the different ratios of the mean difference versus pooled standard deviation; the conditions manipulated in the current study could statistically benefit the standardized version more than the unstandardized counterpart. Finally, also valuable would be further work examining how visible a "medium" effect size is to "careful observers" might also be in order, although this work would likely have to be sophisticated about controlling many different factors (e.g., anchoring effects).

References

1. Cooper H. Synthesizing research: A guide for literature reviews (3rd ed.). Newbury Park, CA: Sage; 1998.
2. Cooper HM, Hedges LV, Valentine JC. The handbook of tesearch synthesis. New York: Russell Sage Foundation; 2009.
3. Glass GV, McGaw B, Smith ML. Meta-analysis in social research. Beverly Hill: Sage Publications; 1981.
4. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando: Academic Press; 1985.
5. Huedo-Medina TB, Johnson BT. Modelos Estadísticos en Meta-análisis [Statistical Models in Meta-analysis]: La Coruña, Spain: Netbiblio; 2010.
6. Hunter JE, Schmidt FL. Methods of meta-analysis: correcting error and bias in research findings. 2nd ed. Thousand Oaks, Calif: Sage; 2004.
7. Johnson BT, Boynton MH. Cumulating evidence about the social animal: Meta-analysis in social-personality psychology. Society & Personality Psychology Compass 2008;2:817-841. doi: 10.1111/j.1751-9004.2007.00048.x
8. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. British Medical Journal. 2001;323(7303):42-46. PMID: 11440947
9. Rosenthal R. Meta-analytic procedures for social research. Rev. ed. Newbury Park: Sage Publications; 1991.
10. Rothstein H, Sutton AJ, Borenstein M. Publication bias in meta-analysis. In H. R. Rothstein, A.J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: prevention, assessment and adjustments. Chichester, UK: Wiley; 2005.
11. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. British Medical Journal. 2000;320(7249):1574-1577. PMID: 10845965
12. Bond CF, Jr., Wiitala WL, Richard FD. Meta-analysis of raw mean differences. Psychological Methods. 2003;8:406-418. PMID: 14664679
13. Blalock HM. Causal inferences in no experimental research. Chapel Hill: University of North Carolina Press; 1964.
14. Pedhauzer EJ. Multiple regression in behavioral research: Explanation and prediction. Fort Worth, TX: Harcourt; 1997.
15. Tukey JW. Analyzing data: Sanctification or detective work? American Psychologist. 1969;24:83-91. doi: 10.1037/h0027108
16. Lipsey MW, Wilson DB. Practical Meta-analysis. Vol 49. Thousand Oaks, CA: Sage Publications; 2001.
17. Glaser RR. Accuracy of ES calculation methods for repeated measures and ANCOVA data. Dissertation. University of Memphis; 2002. ISSN: 0419-4217
18. Ray JW, Shadish WR. How interchangeable are different estimators of effects size? Journal of Consulting and Clinical Psychology. 1996;64:1316-1325. PMID: 8991318
19. Becker BJ. Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology. 1988;41:257-278. doi: 10.1111/j.2044-8317.1988.tb00901.x
20. Cortina JM, Nouri H. ES for ANOVA designs. Thousand Oaks, CA: Sage; 2000.
21. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods. 1996;1:170-177. doi: 10.1037/1082-989X.1.2.170
22. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-group designs. Psychological Methods. 2002;7:105-125. PMID: 11928886

23. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. 2008.

24. Cohen J. Quantitative methods in psychology: A power primer. Psycholical Bulletin. 1992;112(1):155-159. PMID: 19565683

25. Muthen LK, Muthen BO. How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. Structural Equation Modeling. 2002;9(4):22. doi:10.1207/S15328007SEM0904_8

26. McWilliams L. Variance heterogeneity in empirical studies in education and psychology. Paper presented at the annual colloquium of the American Educational Research Association, San Francisco;1991.

27. Wilcox RR. New designs in analysis of variance. Annual Review of Psychology. 1987;38:29-60. doi:10.1146/annurev.ps.38.020187.000333

28. Cohen J. Statistical power analysis for the behavioral sciences (2nd Ed). Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

29. Fleishman AI. A method for simulating nonnormal distributions. Psychometrika. 1978;43:521-531. doi: 10.1007/BF02293811

30. Biggerstaff BJ, Tweedie RL. Incorporating variability estimates of heterogeneity in the random effects model in meta-analysis. Statistics in Medicine. 1997;16:753- 768. PMID: 9131763

31. Brockwell SE, Gordon RI. A comparison of statistical methods for meta-analysis. Statistics in Medicine. 2001;20:825-840. PMID: 11252006

32. Erez A, Bloom MC, Wells MT. Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. Personnel Psychology. 1996;49:275-306. doi: 10.1111/j.1744-6570.1996.tb01801.x

33. Field AP. Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. Psychological Methods. 2001;6:161-180. doi: 10.1037/1082-989X.6.2.161

34. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychological Methods. 1998;3:486-504. doi: 10.1037/1082-989X.3.4.486

35. Overton RC. A comparison of fixed- effects and mixed (random- effects) models for meta-analysis tests of moderator variable effects. Psychological Methods. 1998;3:354-379. doi: 10.1037/1082-989X.3.3.354

36. Shadish WR, Robinson L, Lu C. ES: A computer programs for ES calculation. St Paul, MN: Assesment Systems Corporation; 1999.

37. Gibbons RD, Hedeker DR, Davis JM. Estimation of ES from a series of experiments involving paired comparisons Journal of Educational Statistics. 1993;18:271-279. doi: 10.2307/1165136

38. Hedges LV. Distribution theory for Glass' estimator of effect size and related estimators. Journal of Educational Statistics. 1981;6:107-128. doi: 10.2307/1164588

39 Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1969.

40 Johnson BT, Kirsch I. Do antidepressants work? Statistical significance versus clinical benefits. Significance. 2008;5(2):54-58. doi: 10.1111/j.1740-9713.2008.00286.x