

Meta-Analytic Statistical Inferences for Continuous Measure Outcomes as a Function of Effect Size Metric and Other Assumptions



Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Meta-Analytic Statistical Inferences for Continuous Measure Outcomes as a Function of Effect Size Metric and Other Assumptions

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

Contract No. 290-2007-10067-I

Prepared by:

University of Connecticut–Hartford Hospital Evidence-Based Practice Center

Investigators:

Blair T. Johnson, Ph.D.
Tania B. Huedo-Medina, Ph.D.

This report is based on research conducted by the University of Connecticut, Hartford Hospital Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10067-I). 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 EffectiveHealthCare@ahrq.hhs.gov.

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

Suggested citation: Johnson BT, Huedo-Medina TB. Meta-Analytic Statistical Inferences for Continuous Measure Outcomes as a Function of Effect Size Metric and Other Assumptions. (Prepared by the University of Connecticut, Hartford Hospital Evidence-Based Practice Center under Contract No. 290-2007-10067-I.) AHRQ Publication No. 13-EHC075-EF. Rockville, MD: Agency for Healthcare Research and Quality; April 2013.
www.effectivehealthcare.ahrq.gov/reports/final.cfm.

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 methodological 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 email 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, Ph.D.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge comments from Dr. Parivash Nourjah (TOO), Dr. Issa J. Dahabreh, and the peer reviewers on prior drafts of this report.

Peer Reviewers

Dr. Juan Botella
Facultad de Psicología
Universidad Autónoma de Madrid
Madrid, Spain

Tom D. Stanley, Ph.D.
Department of Economics
Hendrix College
Conway, AR

Dr. Carmen Ortego-Mate
Departamento de Enfermería
Departamento de Ciencias Medicas y
Quirúrgicas
University of Cantabria
Santander, Spain

David B. Wilson, Ph.D.
Criminology, Law and Society
George Mason University
Fairfax, VA

Terri Pigott, Ph.D.
Loyola University Chicago
School of Education
Chicago, IL

Meta-Analytic Statistical Inferences for Continuous Measure Outcomes as a Function of Effect Size Metric and Other Assumptions

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 continuous measures across studies, contemporary meta-analytic practice uses the 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's variance directly in the effect size itself (i.e., $SMD = [M_E - M_C] / SD$) and not directly in the analytical weights. The UMD approach has been conventional even though its bias and efficiency are unknown; these have also not been compared with the SMD. Also unresolved is which of many possible available equations best optimize statistical modeling for the SMD in use with repeated measures designs (one or two groups).

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 ($k = 10, 20, 50, \text{ and } 100$); (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 post-test (ratios: 1:1, 2:1, and 4:1); (4) the post-test mean of each pseudo experimental group to achieve 3 parametric effect sizes ($\delta = 0.25, 0.50, \text{ and } 0.80$); (5) normal versus nonnormal distributions (4 levels); and (6) the between-studies variance ($\tau^2 = 0, 0.04, 0.08, 0.16, \text{ and } 0.32$). For the second issue, (7) the correlation between the two conditions was manipulated ($\rho_{\text{pre-post}} = 0, 0.25, 0.50, \text{ and } 0.75$).

Results and Conclusions. This investigation provides guidance for statistical practice in relation to meta-analysis of studies that compare two groups at one point in time, or that examine repeated measures for one or two groups. Simulations showed that neither standardized or unstandardized effect size indexes had an advantage in terms of bias or 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, the SMD yields better statistical inferences than UMDs in terms of bias and efficiency. Under high skewness and kurtosis, neither metric has a marked advantage. In general, the standardized index presents the least bias under most conditions and is more efficient than the unstandardized index. Finally, the results comparing estimations of the SMD and its variance suggest that some are preferable to others under certain conditions. The current results imply that the choice of effect size metrics, estimators, and sampling variances can have substantial impact on statistical inferences even under such commonly observed circumstances as normal sampling distributions, large numbers of studies, and studies with large samples, and when effects exhibit

heterogeneity. Although using the SMD may make clinical inferences more difficult, use of the SMD does permit inferences about effect size magnitude. The Discussion considers clinical interpretation of results using the SMD and addresses limitations of the current project.

Contents

Introduction	1
Background.....	1
Objectives	2
Orientation to Method.....	3
Significance of Project.....	4
Methods	10
General Design.....	10
Conditions Manipulated.....	11
Results	13
Part One: Unstandardized Versus Standardized Effect Sizes	13
Overview.....	13
Detailed Analysis, Part 1A.....	13
Detailed Analysis, Part 1B.....	19
Part Two: Using the SMD with Repeated Versus Independent Measures	20
Overview.....	20
Detailed Analysis	20
Discussion	29
Choice of Metric When Meta-Analyzing Continuous Measures.....	29
Optimal Estimations of the Standardized Mean Difference Effect Size (and its Sampling Variance).....	32
Limitations and Future Directions	33
References	35
Glossary of Terms	37
Tables	
Table 1. Standardized mean difference ES estimations (and their components) for a one-group repeated-measures design	5
Table 2. Standardized mean difference (SMD) ES estimations (and their components) for two independent groups	6
Table 3. Estimates of sampling variance for the SMD ES in the one-group design with repeated-measures.....	8
Table 4. Estimates of sampling variances for the SMD ES from two-group designs with repeated-measures.....	8
Table 5. Statistics related to the standardized mean difference (SMD) and unstandardized mean difference (UMD) for designs with two independent groups and continuous measures	9
Table 6. Characteristics of simulated datasets	12
Table 7. Bias, efficiency, and 95% CI coverage for the SMD and UMD under three different inferential circumstances.....	20
Table 8. Findings relevant to meta-analytic practice (effect size and variance choice)	30

Figures

Figure 1. Bias of SMD and UMD as a function of asymmetries in the distributions underlying the effect size estimates	14
Figure 2. Bias of SMD and UMD as a function of the between-studies variance of the distribution.....	14
Figure 3. Bias of SMD and UMD as a function of the parametric effect size.....	15
Figure 4. Bias of SMD and UMD as a function of mean sample size.....	15
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 of 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 the heteroskedasticity.....	19
Figure 11. Bias as a function of the correlation parameter of ES indexes for the one-group, within-subjects design	21
Figure 12. Efficiency as a function of the correlation parameter of the ES indexes	22
Figure 13. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes	23
Figure 14. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes.....	24
Figure 15. Bias as a function of the correlation parameter of the ES indexes.....	25
Figure 16. Efficiency as a function of the correlation parameter of the ES indexes	26
Figure 17. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes.....	27
Figure 18. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes.....	28

Appendixes

Appendix A. Bias and Efficiency Results for Standardized and Raw Mean Differences (Specific Aim 1)	
Appendix B. Bias, Efficiency, and Theoretical Variance for All Effect Size Indexes and Their Variances (Specific Aim 2)	

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 fifth and sixth steps as applied to literatures of studies that report outcomes on a single continuous outcome. Thus, dichotomous outcomes are outside the scope of this study, as are literatures of studies for which continuous outcomes are measured on a variety of measures.

Statistical modeling in 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 for each study result in meta-regression and other meta-analytic statistics (see Tables 1 through 5). In contemporary practice, when comparing treatments for trials that use the same continuous 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's variance directly in the effect size (ES) itself (i.e., $SMD = [M_E - M_C] / SD$; e.g., see equation 6, Table 2) and not directly in the analytical weights. In effect, a UMD approach to meta-analysis (see equation 21, Table 5) more heavily weights individual studies' differences to the extent that they have smaller observed variances and larger samples of observations. A SMD approach to meta-analysis more heavily weights studies' differences to the extent that they have larger samples (equation 22, Table 5); the pooled standard deviation observed for each study is used to create the standardized difference between conditions (equation 6, Table 2). The UMD approach has been conventional even though its bias and efficiency are unknown and have not been compared with those of the SMD. Also unresolved is which of many possible available equations best optimize statistical modeling for the UMD and SMD (Tables 3 and 4).

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, the SMD assumes that the observations are dependent, and while some extant meta-analytic procedures account for this dependency, many others do not (Tables 3 and 4); 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 produce discordant results (Tables 3 and 4).^{12,13}

In summary, it is unknown how much bias appears in the weighted effect sizes and moderator 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 equivalent 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 (SMD) under a wide range of analytic circumstances.

In groups of studies for which a phenomenon is assessed using the same measure in every study, meta-analysts have the choice of examining either standardized effect sizes or leaving study outcomes in the original, unstandardized, measure.¹⁴ For example, blood pressure is always assessed in metric units (usually mmHg, or millimeters of mercury) and meta-analyses of blood pressure outcomes routinely leave it in these units, showing, say, that aerobic exercise lowers systolic blood pressure an average of 6 mmHg relative to controls. Efficacy in antidepressant 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 unstandardized units in order to facilitate comparisons between studies.^{15,16}

Nonetheless, the assumptions underlying such advice must be evaluated. For example, they had primary-level studies in mind rather than comparisons of the results of independent studies, such as is the case in meta-analysis. One issue has to do with unequal variances across studies. 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. For example, antidepressant trials focusing on very severely depressed individuals (e.g., M HAM-D=33) will typically have much larger standard deviations than trials that focus on moderately depressed individuals (e.g., M HAM-D=17). 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. Similarly, parametric inferential statistics, the most developed and used methods in

meta-analysis, routinely must meet the normality assumption (lack of skewness and kurtosis). For an example, see Pedhauzer, 1997.¹⁷

Weights for unstandardized outcomes in meta-analysis routinely use the sample size and the variance (see Table 5),^{14,18} but it is unclear whether meta-analytic inferences will be equivalent for the two solutions. To date, no research has examined the comparability of statistical inferences between the UMD and SMD. In the current work, we consider the case of a design that compares two independent groups such as an experimental group and a control group.

The second objective of the current project is:

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

Statistical modeling in meta-analysis cannot proceed unless each study outcome is on the same metric and an appropriate sampling variance is calculated. As Tables 1 and 2 show, 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-designs, 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 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. To the extent that the ES deviates from the null value, it reflects a greater difference between the groups. In its repeated-measures or within-subjects form, the ES 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. Similarly, there are numerous methods of calculating the ES when the outcome is continuous and their available variances are known to vary.^{12,13} Further, the literature leaves unclear whether the different methods of transformation to standardized mean difference from different statistical information types are equivalent and perform well under different parametric conditions.¹⁹⁻²²

Orientation to Method

For both specific aims, Monte Carlo simulation studies are used to generate 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 (1) evaluate the differences between using unstandardized versus standardized metric of effect size (objective 1); and (2) 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). The simulations gauge the performance of these methods of estimation for both objectives.

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 the variance of the measures, the results of this investigation are of great interest. If meta-analytic statistics and inferences do depend on choice of unstandardized vs. standardized effect sizes under some circumstances, then the findings may have far-ranging implications for the practice of meta-analysis. Moreover, it is also important to know the best estimates of within-subjects ESs (in single- and in two-group designs) and to determine which estimates of variance are best 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 potential to improve the accuracy and progress of science and public health. An overarching goal is to enable more accurate empirical generalizations.

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

No.	Source	Equation	Components
1.	Glass et al. (1981) ¹¹	$d_{ira} = t_d \sqrt{\frac{1}{n} 2(1 - r_{Pre,Post})}$	$t_d = \frac{\bar{Y}_{Diff}}{\sqrt{\frac{2S_{Diff}^2(1 - r_{Pre,Post})}{n}}} = \frac{\bar{Y}_{Pre} - \bar{Y}_{Post}}{S_{Diff} \sqrt{\frac{2(1 - r_{Pre,Post})}{n}}}$ $S_{Diff} = \sqrt{\frac{\sum_{i=1}^n (Y_i^{Diff} - \bar{Y}_{Diff})^2}{n-1}} = \sqrt{S_{Pre}^2 + S_{Post}^2 - 2r_{Pre,Post} S_{Pre} S_{Post}}$ <p>SDiff = Standard deviation of the difference assuming unequal variances. n = number of observations \bar{Y}_{Pre} = pretest mean of measure Y. \bar{Y}_{Post} = post-test mean of measure Y. $r_{Pre,Post}$ = correlation between Y_{Pre} and Y_{Post}.</p>
2.	Rosenthal (1991) ³	$d_{tch} = t_d \sqrt{\frac{1}{n}}$	$t_d = \frac{\bar{Y}_{Diff}}{S_{Diff}} \sqrt{n}$
3.	Becker (1988) ¹⁹	$d_b = c(n-1) \frac{\bar{Y}_{Post} - \bar{Y}_{Pre}}{S_{Pre}}$	$c(n-1) = 1 - \frac{3}{4(n-1) - 1}$ <p>S_{Pre} = standard deviation of the pretest</p>
4.	Gibbons et al. (1993) ²³	$d_g = c(n-1) \frac{\bar{Y}_{Diff}}{S_{Diff}}$	
5.	Huedo-Medina and Johnson (2011) ²⁴	$d_{hw} = c(n-1) \frac{\bar{Y}_{Post} - \bar{Y}_{Pre}}{S_{within-pool}}$	$S_{Within-pool} = \sqrt{\frac{(n-1)S_{Pre}^2 + (n-1)S_{Post}^2}{n-1}} = \sqrt{S_{Pre}^2 + S_{Post}^2}$

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

No.	Source	Equation	Components
6.	Hedges (1981) ²⁵	$d_{hb} = c(N - 2) \frac{\bar{Y}_{Post}^E - \bar{Y}_{Post}^C}{S_{Pooled}}$	$c(N - 2) = 1 - \frac{3}{4(N - 2) - 1}$ $S_{Pooled} = \sqrt{\frac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}}$ <p> $N = n_E + n_C$ S_E = post-test standard deviation of the experimental group S_C = post-test standard deviation of the control group \bar{Y}_{Post}^E = post-test mean of the experimental group \bar{Y}_{Post}^C = post-test mean of the control group </p>
7.	Becker (1988) ¹⁹	$d_b = c(N - 2) \left[\frac{\bar{Y}_{Post}^E - \bar{Y}_{Pre}^E}{S_{Pre}^E} - \frac{\bar{Y}_{Post}^C - \bar{Y}_{Pre}^C}{S_{Pre}^C} \right]$	
8.	Gibbons et al. (1993) ²³	$d_g = c(N - 2) \left[\frac{\bar{Y}_{Diff}^E}{S_{Diff}^E} - \frac{\bar{Y}_{Diff}^C}{S_{Diff}^C} \right]$	
9.	Huedo-Medina and Johnson (2011) ²⁴	$d_{hw} = c(N - 2) \left[\frac{\bar{Y}_{Post}^E - \bar{Y}_{Pre}^E}{S_{within-pool}^E} - \frac{\bar{Y}_{Post}^C - \bar{Y}_{Pre}^C}{S_{within-pool}^C} \right]$	

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

No.	Source	Equation	Components
10.	Shadish et al. (1999) ²⁶	$d_{s1} = \frac{\bar{Y}_{Post}^E - \bar{Y}_{Post}^C}{S_{Pooled}}$	$S_{Pooled} = \sqrt{\frac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}}$ $S_E = \frac{ \bar{Y}_{Pre}^E - \bar{Y}_{Post}^E \sqrt{n_t}}{ t_d^E \sqrt{2(1 - r_{Pre,Post}^E)}}$ $S_C = \frac{ \bar{Y}_{Pre}^C - \bar{Y}_{Post}^C \sqrt{n_t}}{ t_d^C \sqrt{2(1 - r_{Pre,Post}^C)}}$
11.	Shadish et al. (1999) ²⁶	$d_{s2} = \frac{\bar{Y}_{Post}^E - \bar{Y}_{Post}^C}{S_{ANOVA}}$	$S_{ANOVA} = \sqrt{\frac{MSE_b + (tp - 1)MSE_w}{tp}}$ <p>MSE_b = between-subjects mean square error MSE_w = within-subjects mean square error tp = number of measured time points</p>
12.	Shadish et al. (1999) ²⁶	$d_{s3} = \frac{\bar{Y}_{Post}^E - \bar{Y}_{Post}^C}{S_{ANCOVA}}$	$S_{ANCOVA} = \sqrt{\frac{MSE_a(N - h - 1)}{(1 - r_{w-class}^2)(N - h)}}$ $r_{w-class} = \sqrt{\frac{F_{cov}}{F_{cov} + (N - h - 1)}}$ <p>MSE_a = adjusted mean square error from the covariance analysis</p>

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	$\text{var}_{\text{one-g}}(d_b) = \left(\frac{2(1-r_{pre,post})}{n} \right) \left(\frac{n-1}{n-3} \right) \left(1 + \frac{n}{2(1-r_{Pre,Post})} d_b^2 \right) - \frac{d_b^2}{[c(100-1)]^2}$
14.	Change-score metric (Gibbons et al., 1993) ²³	$\text{var}_{\text{one-g}}(d_g) = \left(\frac{1}{n} \right) \left(\frac{n-1}{n-3} \right) (1 + nd_g^2) - \frac{d_g^2}{[c(n-1)]^2}$

Note: “Raw-score” implies having the measures’ variability in the original 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) ²⁵	$\text{var}_{\text{two-g}}(d_{b-t}) = \left(\frac{1}{\tilde{n}} \right) \left(\frac{N-2}{N-4} \right) (1 + \tilde{n}d_{b-t}^2) - \frac{d_{b-t}^2}{[c(N-2)]^2}, \quad \tilde{n} = \frac{n_E * n_C}{n_E + n_C}$
16.	Raw-score metric for a function of two ESs (Becker, 1988) ¹⁹	$\text{var}_{\text{two-g}}(d_b) = \text{var}_{\text{one-g}}(d_b^E) + \text{var}_{\text{one-g}}(d_b^C)$
17.	Change-score metric for a total ES (Morris & DeShon, 2002) ²²	$\text{var}_{\text{two-g}}(d_{g-t}) = \left(\frac{1}{2(1-r_{Pre,Post}^E)\tilde{n}} \right) \left(\frac{N-2}{N-4} \right) (1 + 2(1-r_{Pre,Post}^E)\tilde{n}d_{g-t}^2) - \frac{d_{g-t}^2}{[c(N-2)]^2}$
18.	Change-score metric for a function of two ESs (Morris & DeShon, 2002) ²²	$\text{var}_{\text{two-g}}(d_g) = \text{var}_{\text{one-g}}(d_g^E) + \text{var}_{\text{one-g}}(d_g^C)$

Note: “Raw-score” implies having the measures’ variability in the original observations and does not imply the unstandardized mean difference (UMD).

Table 5. Statistics related to the standardized mean difference (SMD) and unstandardized mean difference (UMD) for designs with two independent groups and continuous measures

No.	Metric	Equation
19.	Unstandardized mean difference (UMD) for two independent groups ¹⁸	$UMD = \bar{Y}_{Post}^E - \bar{Y}_{Post}^C$
20.	Standard error of the UMD ¹⁸	$SE_{UMD} = S_{Pooled} \sqrt{\frac{1}{n_E} + \frac{1}{n_C}}$
21.	Inverse variance of the UMD ¹⁸	$\frac{1}{Var_{UMD}} = \frac{n_E n_C}{S_{Pooled}^2 (n_E + n_C)}$
22.	Inverse variance of the SMD ¹⁸	$\frac{1}{Var_{SMD}} = \frac{2n_E n_C (n_E + n_C)}{2(n_E + n_C)^2 + n_E n_C SMD^2}$
23.	Standard error of the SMD ¹⁸	$SE_{SMD} = \sqrt{\frac{n_E + n_C}{n_E n_C} + \frac{SMD^2}{2(n_E + n_C)}}$

Methods

General Design

- 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 the R command `mvnrm`. (Both of these R commands rely on the Marsenne-Twister²⁷ random number generator.) 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.
- We generated two bivariate normal distributions, each with a homogeneous variance-covariance matrix,

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

representing the experimental and control groups, respectively; only the Y^E matrix was generated in the case of single-group designs. The parameters for these distributions in the standardized units are $\mu_{Pre}^E = \mu_{Post}^C = \mu_{Pre}^C = 0$, with $\sigma_{Pre,Post}$, σ_{Pre}^2 , σ_{Post}^2 , and μ_{Post}^E 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 using 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 bias of the estimate was calculated:

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

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,

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

- 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 with the average of the variance obtained from each sampling variance estimate to obtain the adjustment to the theoretical variance.

Conditions Manipulated

Monte Carlo simulations established many different conditions by crossing these factors (see Table 6) to evaluate Specific Aim 1:

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 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 post-test measures were varied using ratios for experimental and control groups, respectively, of 1:1, 2:1, and 4:1.^{28,29} 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.¹¹
4. The mean of the post-test for the experimental group, following the parametric values for the standardized mean difference,³⁰ $\delta = \mu_{post}^E = 0.25, 0.5,$ and 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 nonnormal 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 nonnormal distributions, the normality pattern was manipulated to obtain skewed distributions through use of the Fleishman³¹ 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 $\tau^2 = 0$, statistical models reduce 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.^{32,33}

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 $\rho_{\text{pre-post}} = 0, 0.25, 0.50, \text{ and } 0.75$.

Table 6. Characteristics of simulated datasets

Condition	Levels
Specific Aim 1	
Mean population effect size	$\delta = \mu_{\text{post}}^E = 0.25 / 0.50 / 0.80$
Relative within-study standard deviation for control and experimental groups	$\sigma_E : \sigma_C = 1:1, 2:1, \text{ and } 4:1$
Sample size vectors	$n_C = n_E = [12, 16, 18, 20, 84], [32, 36, 38, 40, 104], \text{ and } [62, 66, 68, 70, 134]$
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	$\tau^2 = 0 / 0.04 / 0.08 / 0.16 / 0.32$
Only for Specific Aim 2	
Correlation between pretest and post-test measures	$\rho_{\text{pre-post}} = 0 / 0.25 / 0.50 / 0.75$

Results

Part One: Unstandardized Versus Standardized Effect Sizes

Overview

Monte Carlo simulations (Part 1A) 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. A further simulation (Part 2B) showed that the two indexes converge in terms of bias, efficiency, and coverage when data are normally distributed and studies are homogeneous in gauging the parametric ES.

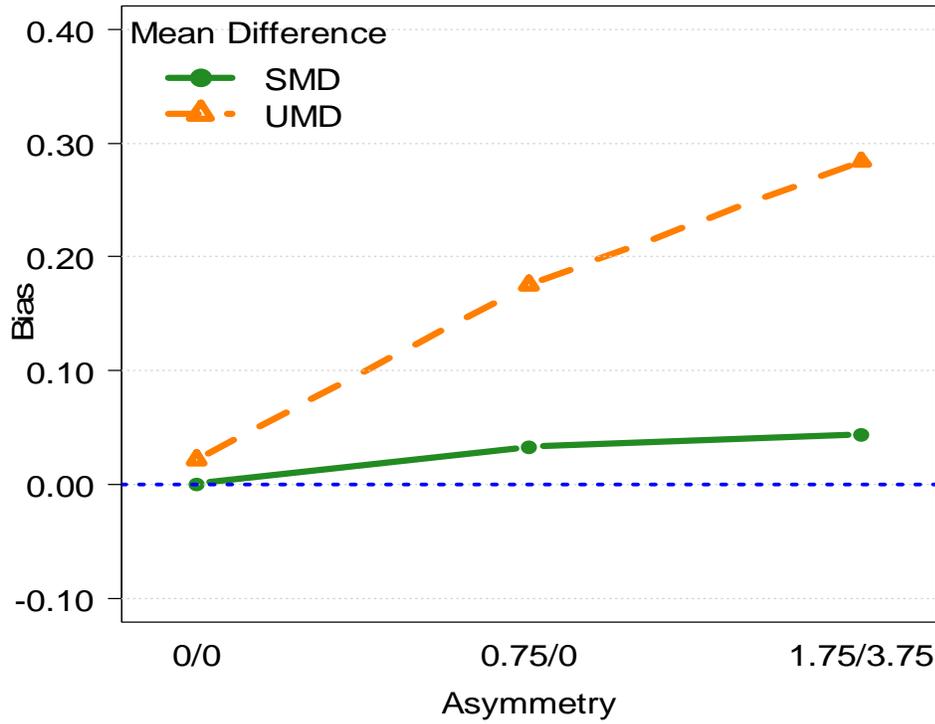
Detailed Analysis, Part 1A

In this section we detail the initial Monte Carlo simulation study that we performed. In Detailed Analysis, Part 2, we describe a second simulation study that addressed issues that emerged from the first.

Bias

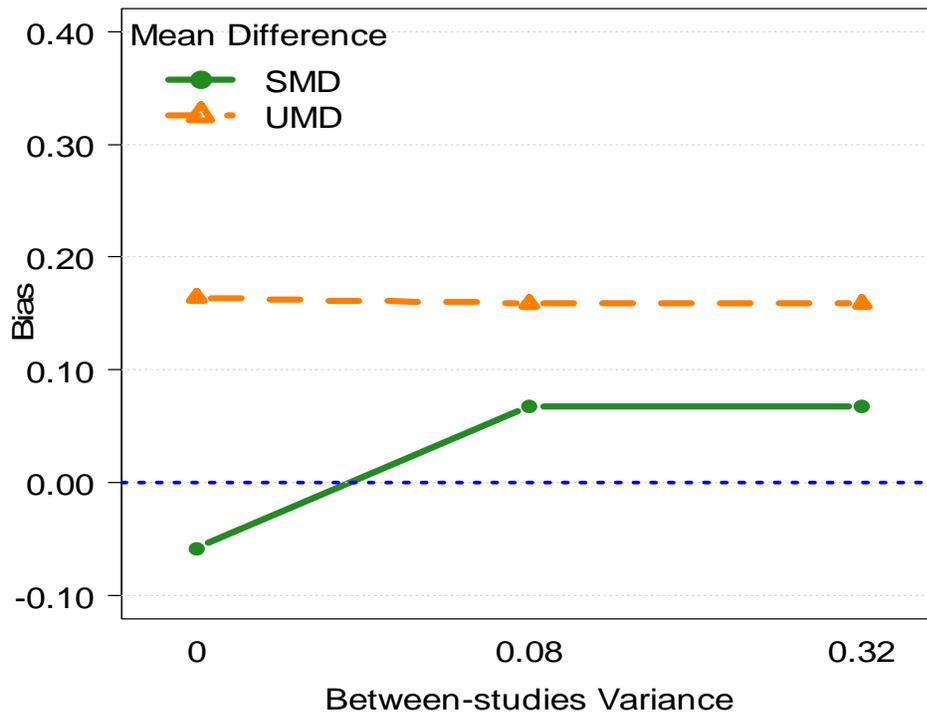
Figures 1 through 5 show the primary results comparing the UMD against the SMD in terms of bias. As these figures show, the SMD had less bias than did the UMD under all plotted circumstances. The two metrics approach the same level of bias only when skewness and kurtosis is minimal (Figure 1), but even here the SMD showed a slight advantage. More dramatic differences between the two appeared as skewness and kurtosis increase, under fixed-effects ($\tau^2=0$; see Figure 2), as the parametric effect size increases (Figure 3), as the mean sample size increases (Figure 4), and as the difference in variances between the two compared groups 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 overestimate the parametric effect size.

Figure 1. Bias of SMD and UMD as a function of asymmetries in the distributions underlying the effect size estimates



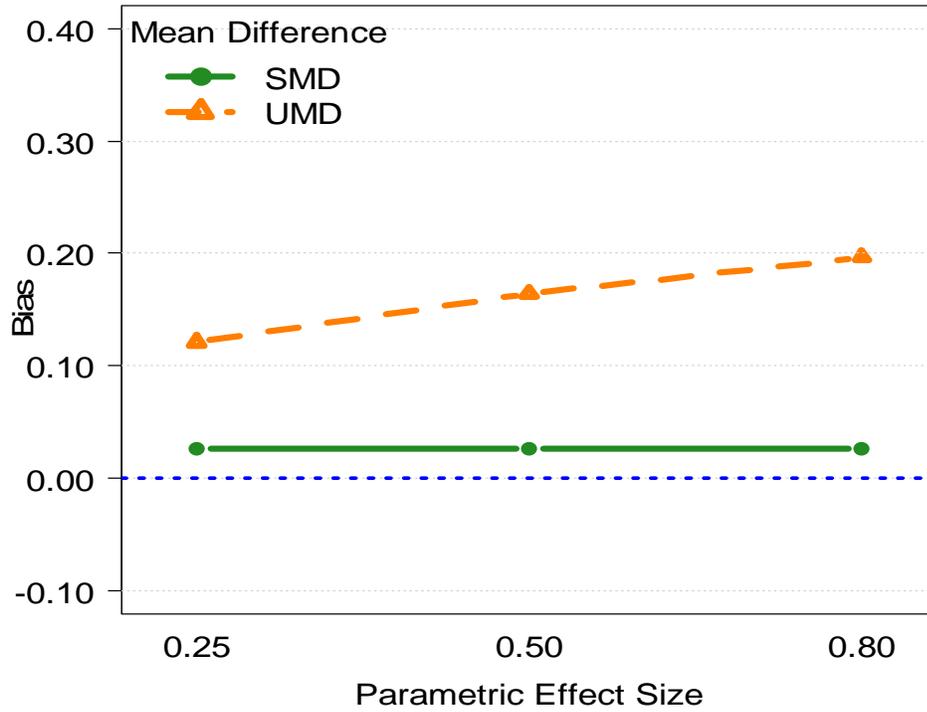
SMD = standardized mean difference; UMD = unstandardized mean difference

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



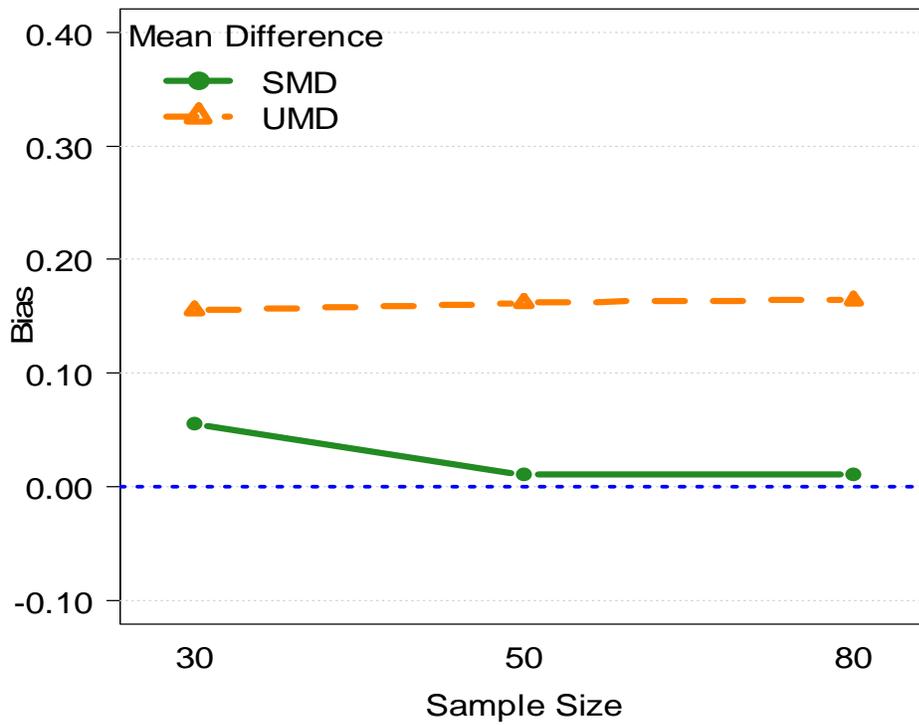
SMD = standardized mean difference; UMD = unstandardized mean difference

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



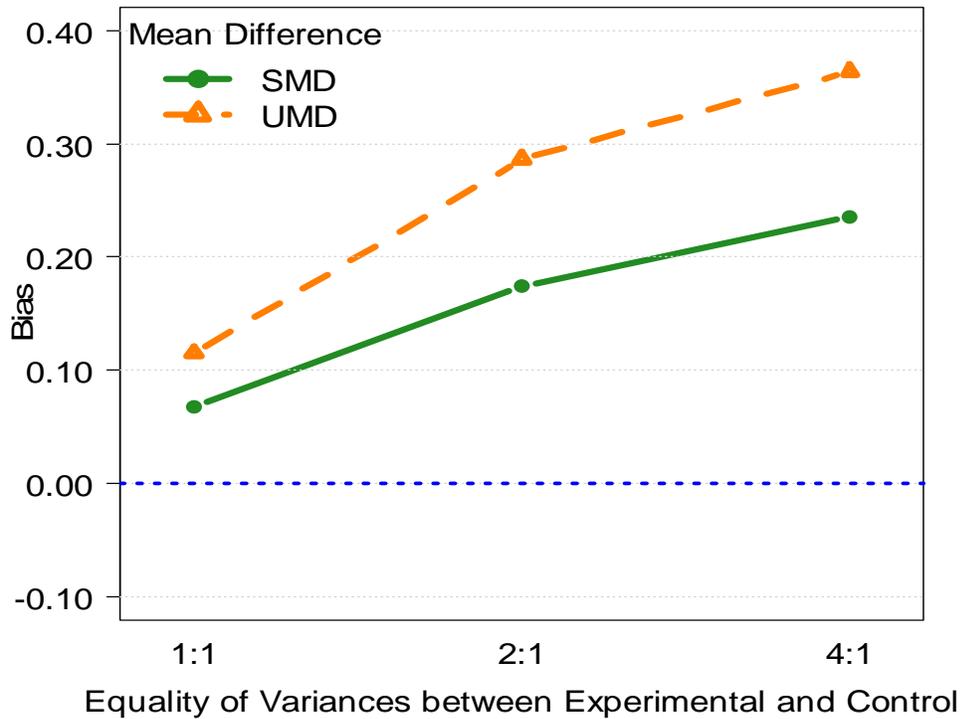
SMD = standardized mean difference; UMD = unstandardized mean difference

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



SMD = standardized mean difference; UMD = unstandardized mean difference

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

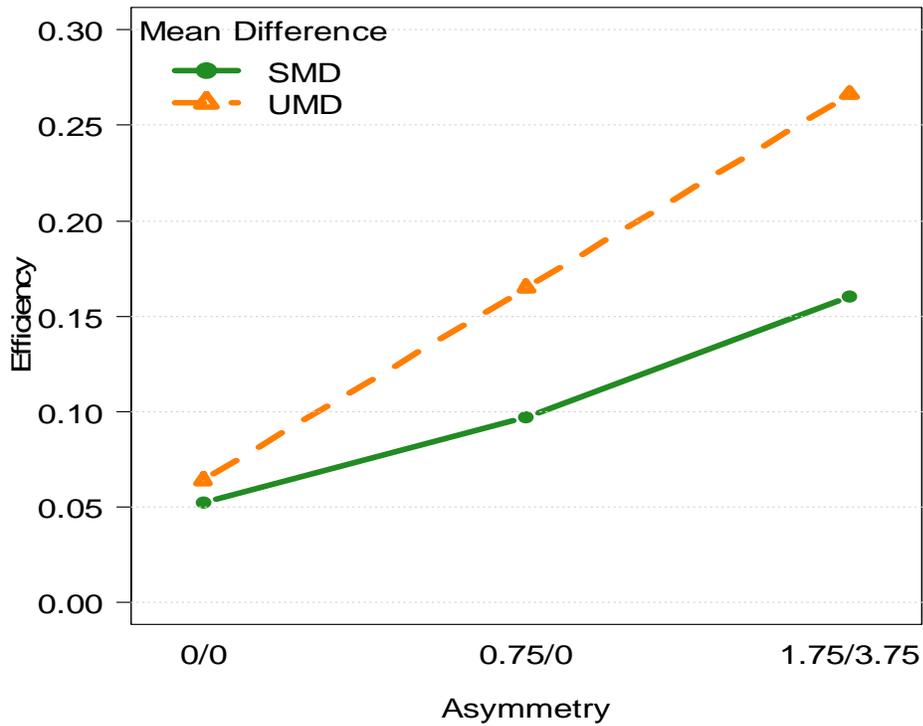


SMD = standardized mean difference; UMD = unstandardized mean difference

Efficiency

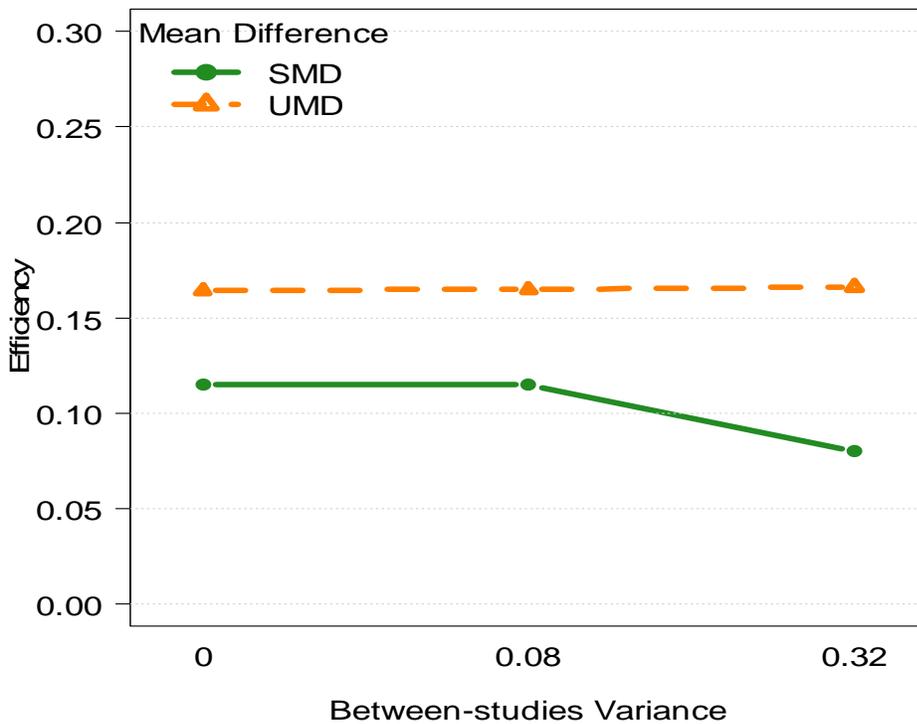
Figures 6 through 10 show the primary results comparing the UMD against the SMD in terms of efficiency, such that values nearer to 0 are more efficient (i.e., values close to zero reflect less variability, so more efficiency to detect a difference between the compared groups' means). The SMD presented more efficient estimations than the UMD under the same circumstances as it exhibited less efficiency. 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 the difference between the variances of the two groups increases (Figure 10). 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 (and these are not plotted in the main report; see Appendix A, Tables A1 to A8 for more detailed results).

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



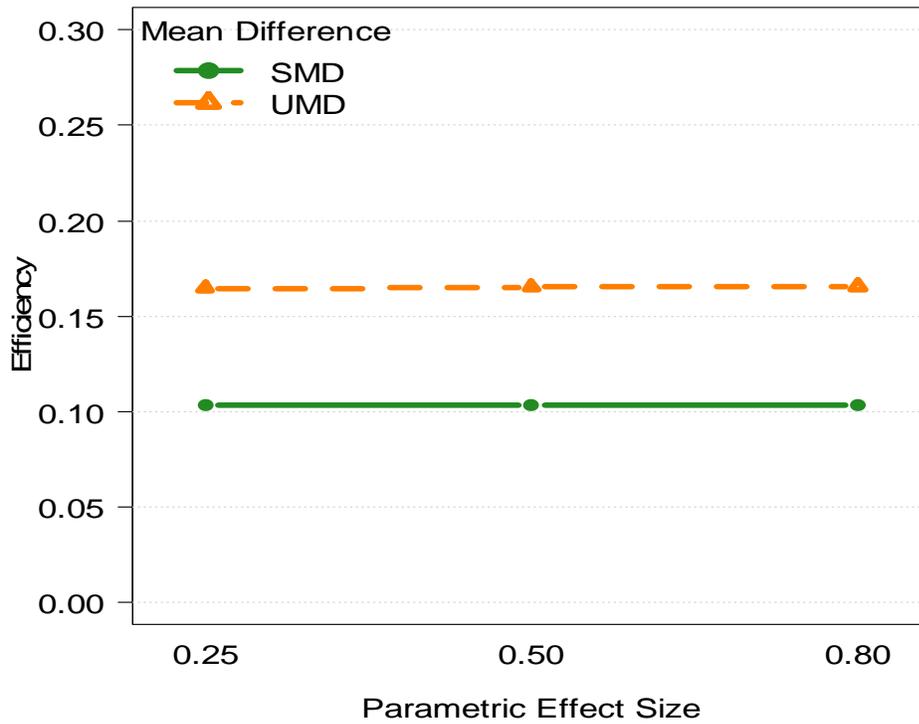
SMD = standardized mean difference; UMD = unstandardized mean difference

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



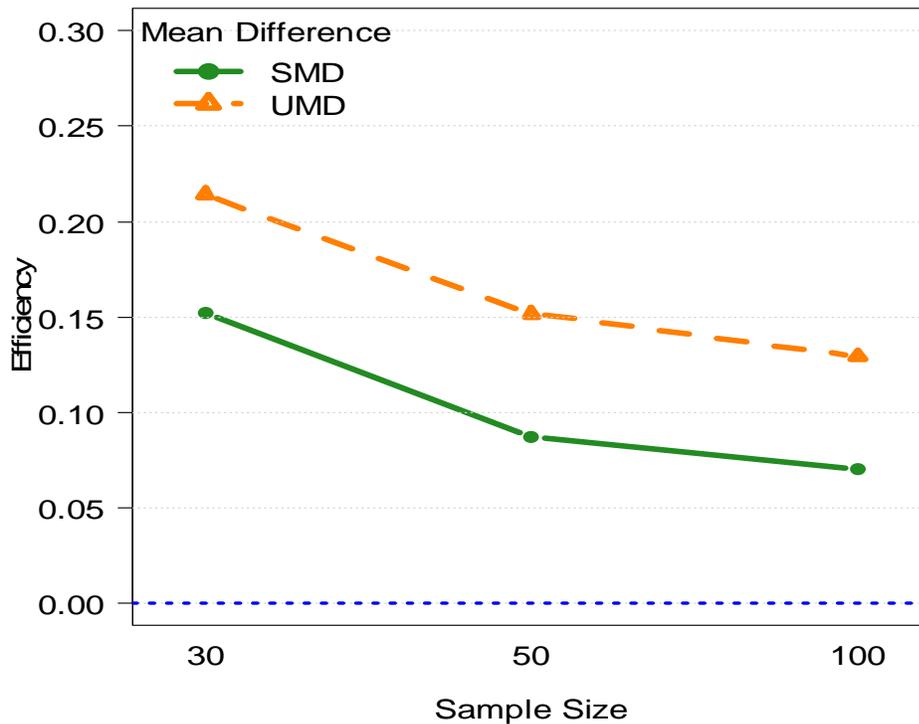
SMD = standardized mean difference; UMD = unstandardized mean difference

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



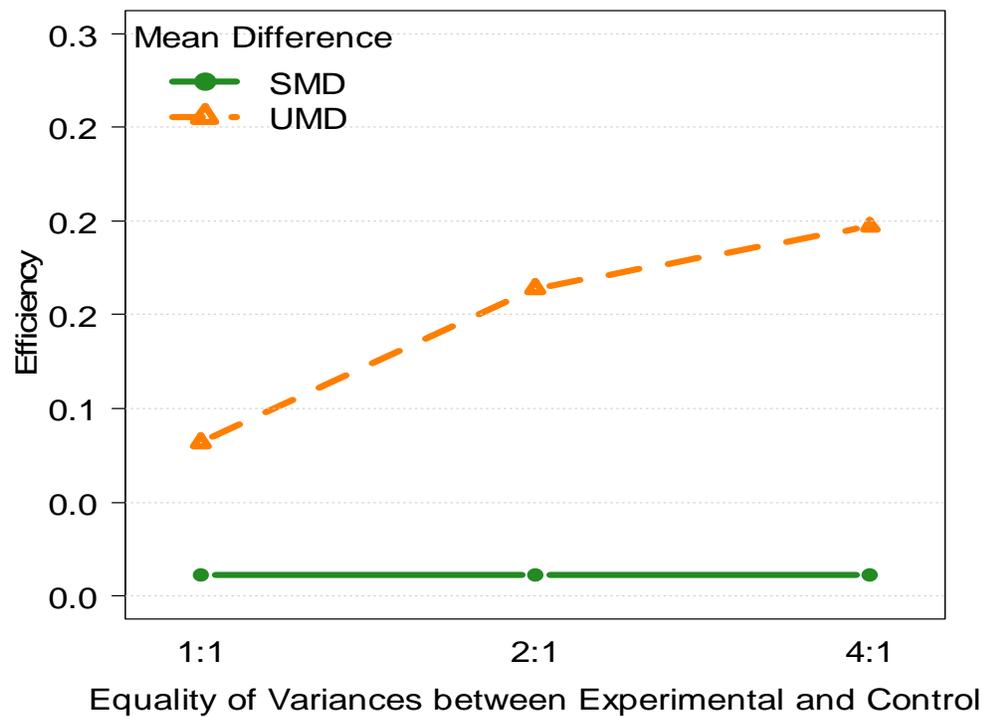
SMD = standardized mean difference; UMD = unstandardized mean difference

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



SMD = standardized mean difference; UMD = unstandardized mean difference

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



SMD = standardized mean difference; UMD = unstandardized mean difference

Detailed Analysis, Part 1B

We followed the same procedures in a subsequent Monte Carlo simulation study in order to explore the performance of the two indexes, SMD and UMD, when a weighted mean effect size is obtained under particular conditions and evaluating the 95% confidence interval (CI) coverage; the previous simulation evaluated only whether individual studies approximated the parametric effect size. We envisioned circumstances that might be regarded as “well-behaved data” compared with “unruly” data, where both types of data gauged the same parametric medium effect size ($\delta=0.50$). For our well-behaved data, we simulated normal distributions with no skewness or kurtosis, large sample sizes ($Mn = 80$), and a large number of studies ($k = 100$). We conceived of two unruly data conditions with the same parameters, except that in one the experimental vs. control variances very unequal (4:1), and in the other there was an extremely nonnormal distribution (skewness/kurtosis = 1.75/3.75). There were 10,000 replications in each condition. Note that these conditions do not evaluate random-effects circumstances (i.e., $\tau^2=0$ in this simulation).

Bias and efficiency were obtained across the replications as in Part 1, but coverage of the confidence interval also was obtained specially for this set of simulations. Thus, the final average weighted effect size across the individual studies was generated along with a 95% confidence interval around the mean effect size. Then the coverage of their confidence intervals was obtained as the proportion of replications in which the confidence interval for each index did not include the null value, $ES = SMD = UMD = 0$.

In obtaining the weighted ES, the SMD and UMD indexes exhibited bias and efficiency that were only trivially different (Table 7). Both slightly overestimated the parametric ES under either the normal data circumstance or the nonnormal distributions, and both dramatically

overestimated the ES under circumstances of unequal variances. These results serve as a form of replication of the results of Part 1A.

The coverage of their CIs indicates that both the SMD and UMD estimations include the true value in 95 percent of the simulations. In some occasions, both CIs missed the parameter under unequal variances or when nonnormal distributions are present.

Table 7. Bias, efficiency, and 95% CI coverage for the SMD and UMD under three different inferential circumstances

Statistic	Normal Data, Equal Variances, Large Samples	Unequal Variances (4:1)	Skewness/Kurtosis (1.75/3.75)
<i>Bias</i>			
SMD	0.0169	0.3153	0.0256
UMD	0.0170	0.3250	0.0572
<i>Efficiency</i>			
SMD	0.0824	0.1055	0.1564
UMD	0.0852	0.1047	0.1587
<i>95% CI Coverage</i>			
SMD	0.9521	0.9328	0.9146
UMD	0.9567	0.9339	0.9197

CI = confidence interval; SMD = standardized mean difference; UMD = unstandardized mean difference

Part Two: Using the SMD with Repeated Versus Independent Measures

Overview

- If one is interested in analyzing the effects of an intervention excluding time-related effects, then it is convenient to consider the raw-score metric. That is, research routinely controls for the stability of observations of particular cases across time. If a treatment affects an outcome uniformly across cases, then a perfect correlation between pre- and post-test observations is implied. Treatments that affect the outcome differentially across cases and time imply a correlation that is less than 1. Statistical analyses in primary-level statistics routinely control for this within-subjects variability (see, for example Pedhauzer, 1997¹⁷). In meta-analytic statistics, debate exists about whether to control for this source of variability (see Tables 1–4).
- The Monte Carlo simulation systematically compared estimates for effect size (Tables 1–2) and the sampling variance (Tables 3–4) across studies that varied in the magnitude of the correlation between the pre- and post-tests ($\rho=0, 0.25, 0.50, 0.75$).
- Because the main factor that affected estimates was the magnitude of the correlation between the two measures, other simulated conditions are not presented here (see Appendix B).

Detailed Analysis

One-Group Within-Subjects Design

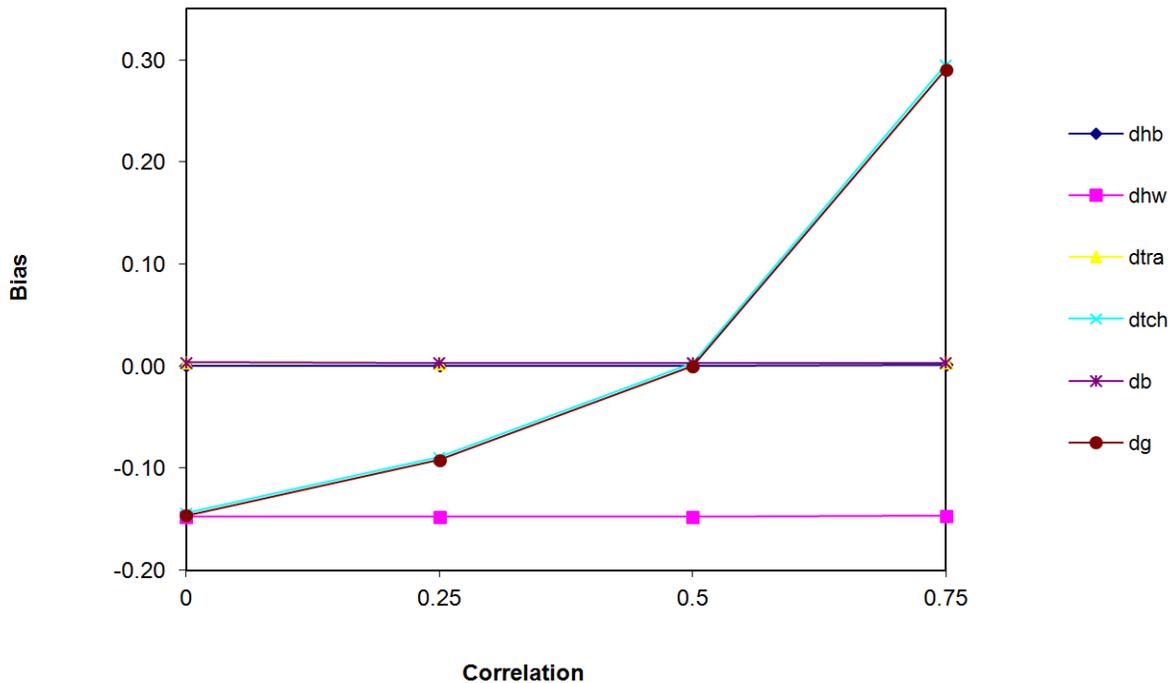
The simulation evaluated bias, efficiency for five different indexes designed for the one-group, within-subjects design, d_{tra} , d_{tch} , d_b , d_g , d_{hw} (see Table 1). The simulation also evaluated d_{hb} (Table 2, equation 6) even though it was designed for a two-group comparison.³⁴ As well as

the theoretical variance adjustment for the variance estimations that have been presented in the literature and combining them across the different indexes.

Bias

Figure 11 shows bias of the six different indexes calculated as the difference between the mean of the two repeated observations and the parametric ES in raw metric, δ_{raw} , over the replications in each correlation condition, assuming normality of distributions. Positive values reflect an overestimation of this particular parameter, whereas negative values imply an underestimation.)

Figure 11. Bias as a function of the correlation parameter of ES indexes for the one-group, within-subjects design



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

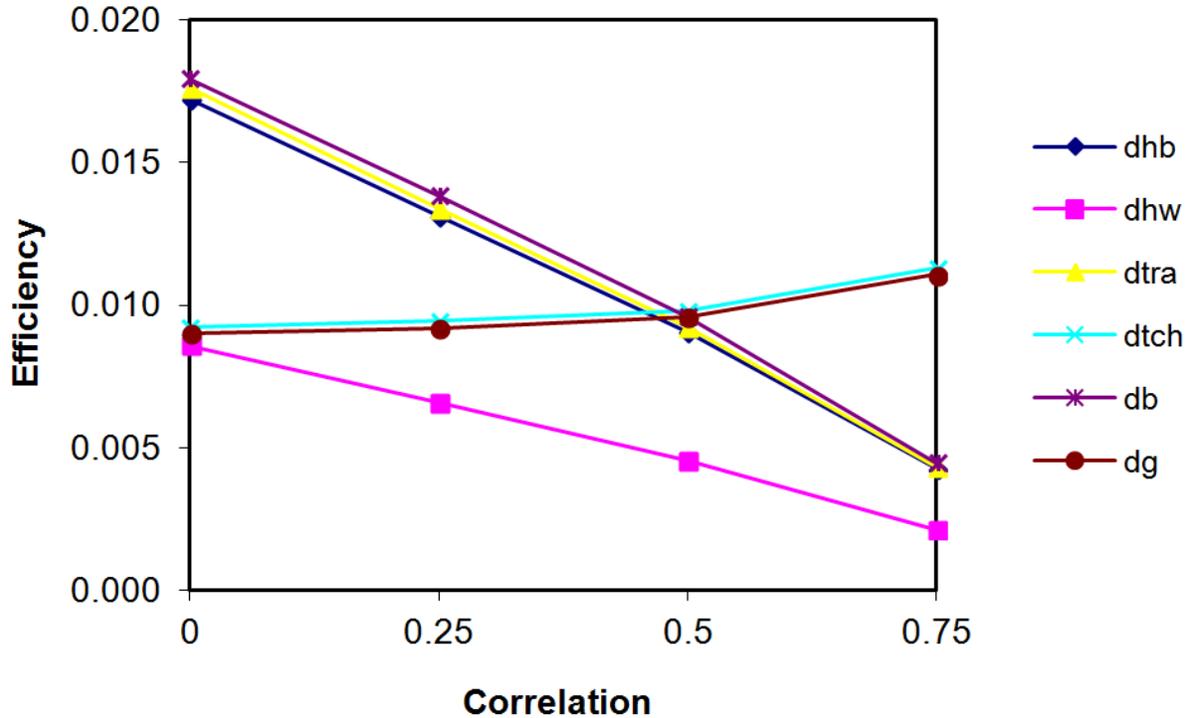
Of all ES indexes, d_b and d_{tra} were the least biased across all correlations and their performance are exactly the same as d_{hb} . Two others had no bias when the correlation was 0.5, d_g and d_{tch} , but underestimated the parametric ES when the correlation was smaller and overestimated it when the correlation was larger. The remaining index in Figure 11 (one that does not explicitly incorporate the correlation in its equation), d_{hw} consistently underestimated the ES regardless of the correlation (see Appendix B, Table B2, for more detail).

Efficiency

As Figure 12 shows, the most efficient ES index was d_{hw} and its efficiency improved as the correlation between the measures grew. Its efficiency was only slightly better than d_{tch} and d_g when $\rho=0$. Moreover, d_b , d_{tra} , and d_{hb} had the lowest efficiency when the repeated measures were less correlated but did better when they were highly correlated. The other two indexes, d_g and

d_{tch} , performed worse under high correlation conditions and better under low correlation conditions. Estimates for d_g and d_{tch} are almost unaffected by increasing the assumed correlation between pre- and post-test, although these increase slightly as the correlation increases. 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.

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

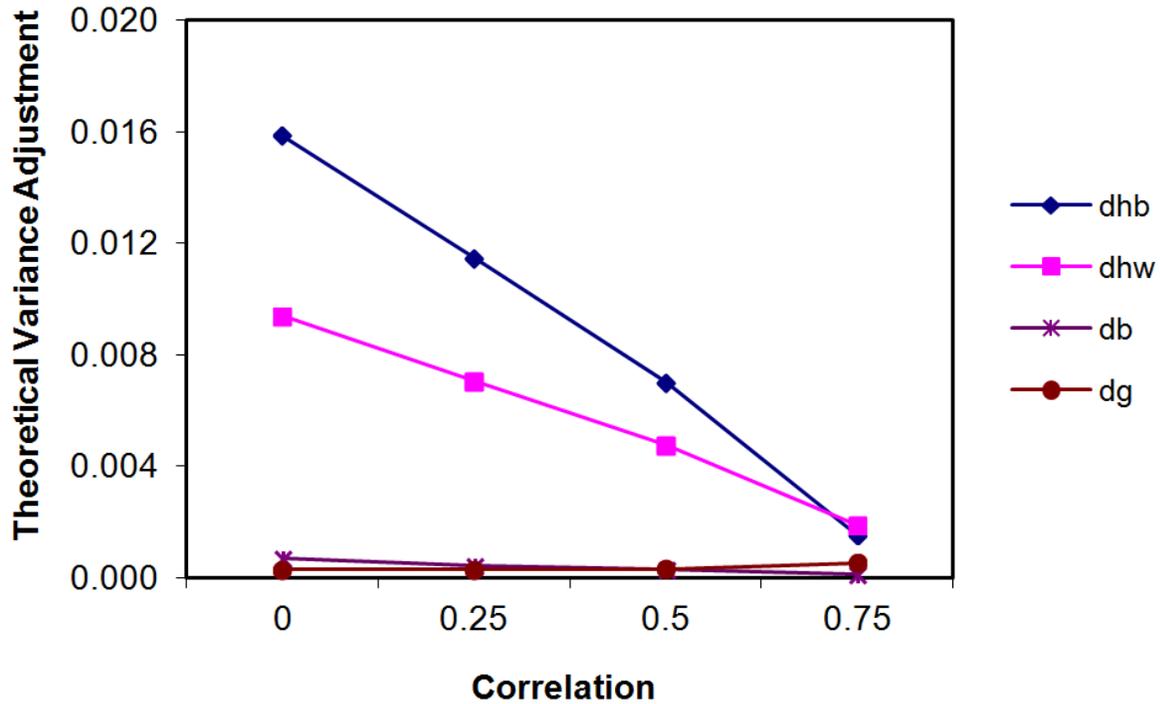


d_{hb} = Hedges's standardized mean difference using the between-studies degrees of freedom; d_{hw} = Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom; d_{tra} = standardized mean difference in raw-score metric form t_d ; d_{tch} = the standardized mean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

The Theoretical Variance Adjustment

Figure 13 shows how the correlation parameter relates to adjustments to the theoretical variance in the two-groups design with repeated measures. Note that here results for d_{hb} , d_{hw} , and d_b used variance equation 13 and d_g used equation 14 (see Table 3). 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.

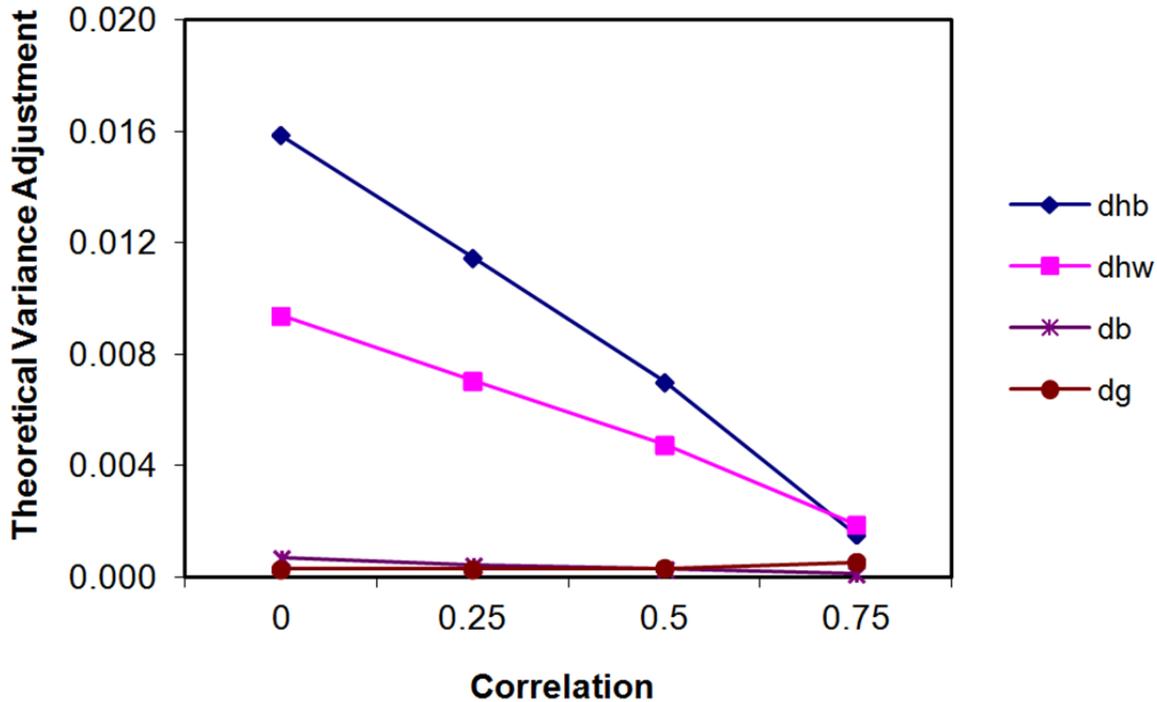
Figure 13. The theoretical variance adjustment to the empirical variance as a function of the correlation parameter of the ES indexes



d_{hb} = Hedges's standardized mean difference using the between-studies degrees of freedom (Equation 13, Table 3);
 d_{hw} = Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom (Equation 13, Table 3); d_b = Becker's standardized mean difference (Equation 13, Table 3); d_g = Gibbons's standardized mean difference (Equation 14, Table 3)

If the theoretical variance is calculated for d_{hw} and d_b without including the correlation factor, $2(1 - \rho)$, and using the $df = n - 1$ and $\tilde{n} = n$, d_{hw_nonr} and d_{b_nonr} , these indexes present opposite patterns to their corresponding 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



d_b = Becker's standardized mean difference; d_{b_nonr} = Becker's standardized mean difference without including the correlation factor $2(1 - r)$ in its equation (Equation 13, Table 3); d_g = Gibbons's standardized mean difference (Equation 14, Table 3); d_{hb} = Hedges's standardized mean difference using the between-studies degrees of freedom (Equation 13, Table 3); d_{hw} = Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom (Equation 13, Table 3); d_{hw_nonr} = Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom without including the correlation factor $2(1 - r)$ in its equation (Equation 14, Table 3)

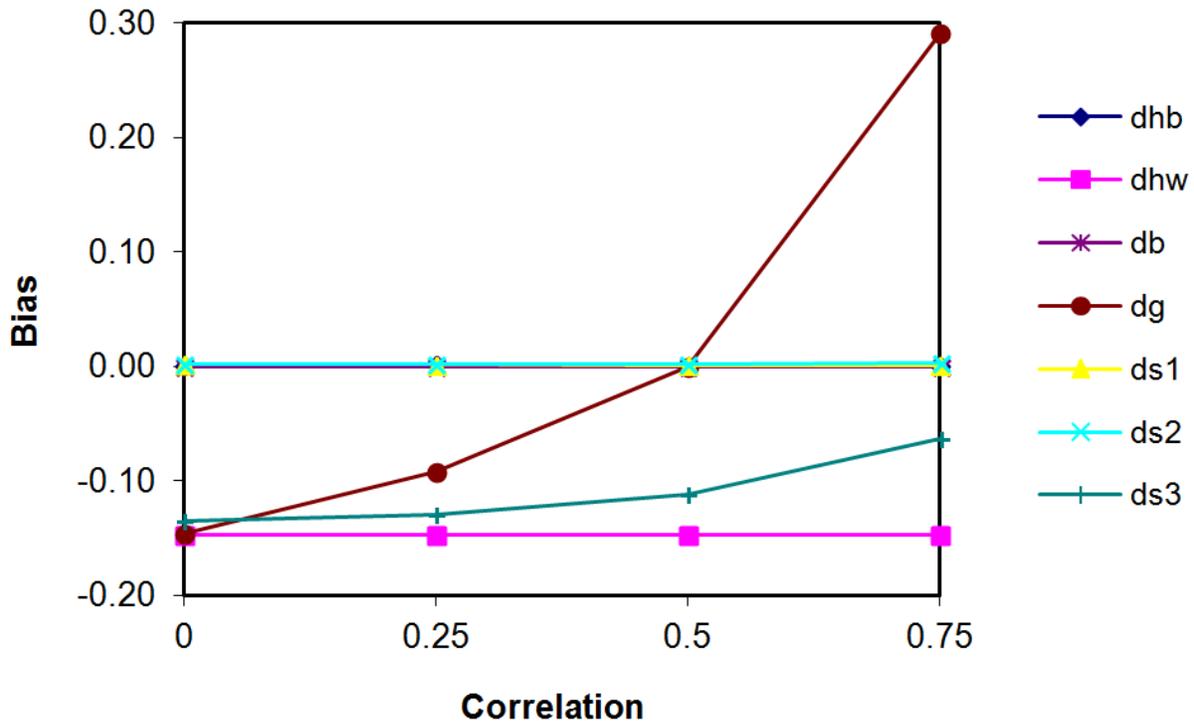
Two-Group Repeated Measures Designs

The simulation evaluated bias, efficiency, and theoretical variance adjustment for seven different estimates of effect sizes (ESs) for the two-groups designs with repeated measures, d_{hb} , d_b , d_g , d_{hw} , d_{s1} , d_{s2} , and d_{s3} (see Table 2). One of these, d_{hb} (Table 2, equation 6) was designed to focus on a two-group comparison at one measurement point (without considering repeated measures); the others utilize an earlier measure in some respect.

Bias

As Figure 15 shows, two ES indexes present no bias in estimating the parametric ES, d_{s1} and d_{s2} . The denominator for each of these equations incorporates the correlation between the two measures in some respect (in the standard deviation or in the mean square error). The other indexes ignore the correlation between the measures and usually exhibit a negative bias. The only exception is that if the measures are correlated moderately ($\rho = 0.5$), d_g performs with no bias; this ES underestimates the parametric effect size when the measures are less correlated and overestimates it when they are more correlated.

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



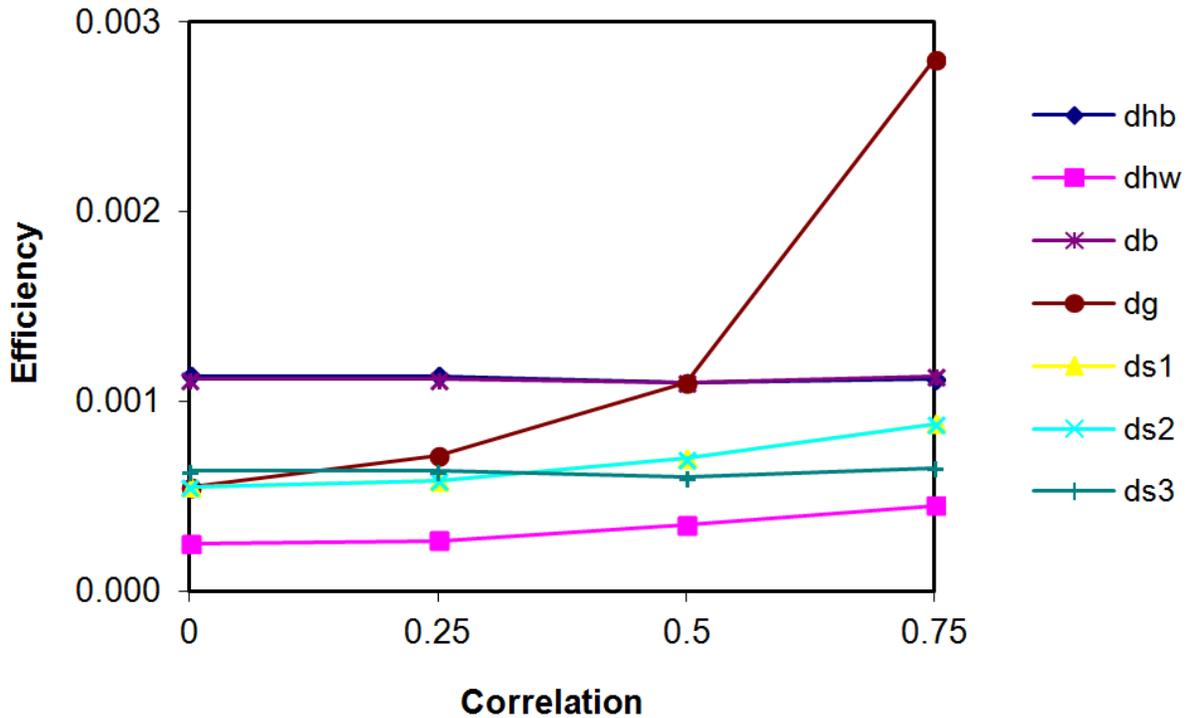
d_{hb} = Hedges's standardized mean difference using the between-studies degrees of freedom; d_{hw} = Huedo-Medina and Johnson's standardized mean difference; d_b = Becker's standardized mean difference; d_g = Gibbons' standardized mean difference; d_{s1} = standardized mean difference using Shadish's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

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 indexes d_{hb} and d_b have very similar efficiency, with d_b being slightly larger when $\rho = 0$; when $\rho = 0.5$, the variability of d_{hb} is slightly larger than d_b (see Figure 16).

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

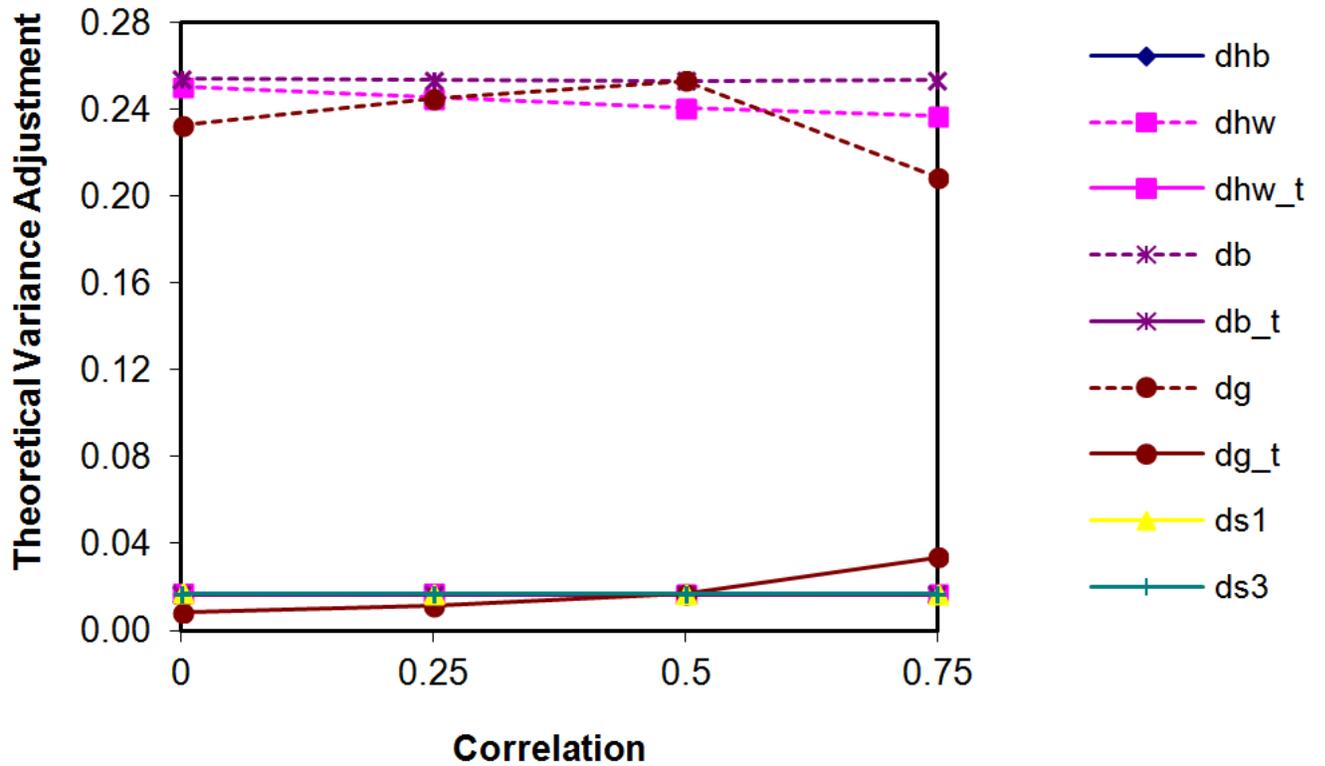


d_{hb} = Hedges's standardized mean difference using the between-studies degrees of freedom; d_{hw} = Huedo-Medina and Johnson's standardized mean difference; d_b = Becker's standardized mean difference; d_g = Gibbons' standardized mean difference; d_{s1} = standardized mean difference using Shadish's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

The Theoretical Variance Adjustment

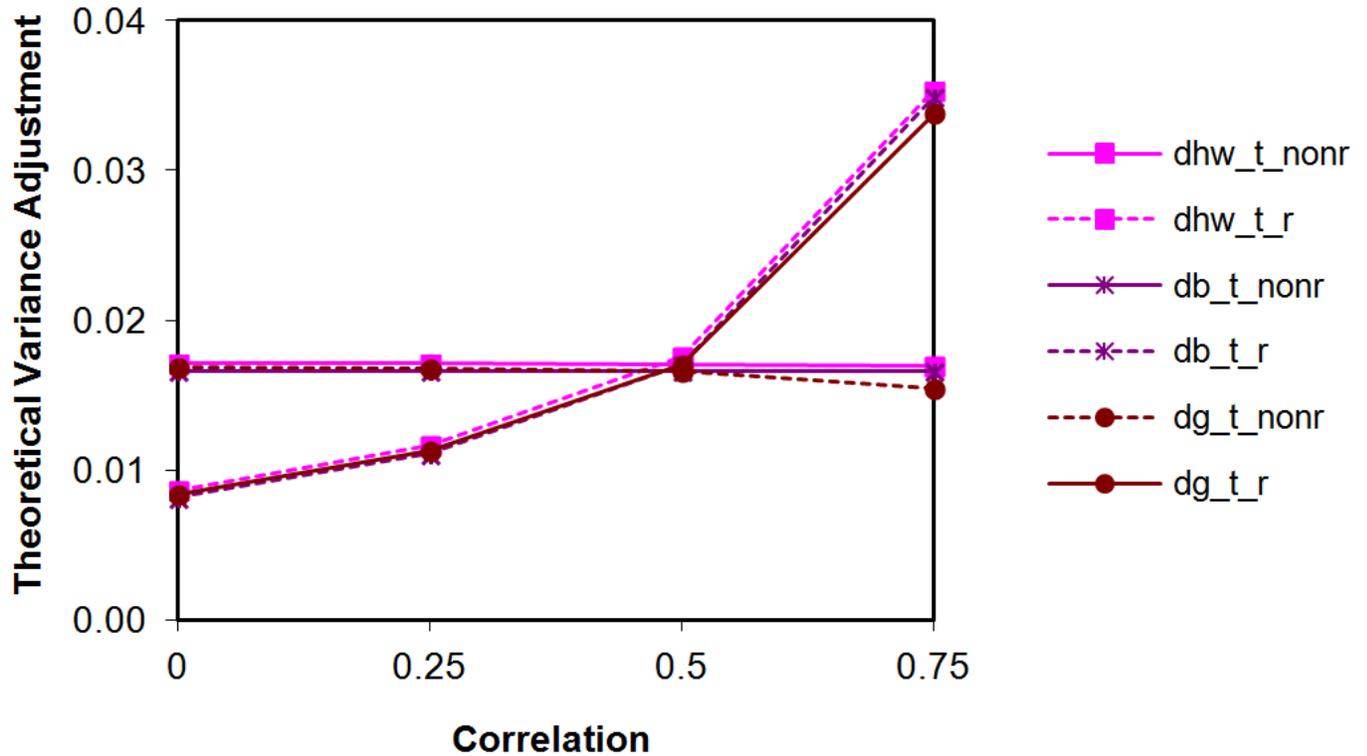
Figures 17 and 18 present the differences between the sampling variance estimates and the actual sampling variance of the effect size index. First, in Figure 17, the estimates illustrated using a solid line are calculated as the variance of an ES, and those dashed lines are obtained as a variance compound of two variances, one for the ES of each independent group. The effect size indexes that require the least theoretical variance adjustment are d_{s1} , d_{s3} , d_g , and d_{hw} , with the latter two effect sizes defined in such a way that the variance calculated for a total ES is calculated as a difference between the ES for each group. The effect size indexes that required the largest theoretical variance adjustment are d_g , d_{hw} , and d_g . The correlation between the measures has relatively little role in affecting the theoretical variance adjustment. (Note that d_{s2} is not considered in Figure 17 because it has exactly the same performance as d_{s1} .) The considerably worse performance appears in those estimates of the sampling variance that are a composite of the experimental and control group variance.

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



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

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



d_{hw} = Hedges's standardized mean difference using the within-study degrees of freedom; d_b = Becker's standardized mean difference; d_g = Gibbons's standardized mean difference; note that the inclusion or exclusion of the correlation factor, $2(1-\rho)$, is denoted as *_r* and *_nonr*, respectively, in the suffix of these terms

Figure 18 shows the sampling variance estimations for a total ES in different metrics, examining those that can include vs. exclude the correlation factor, $2(1 - \rho)$, d_{hw} , d_b , and d_g . As expected, when the correlation factor was not included, all three indexes required the same theoretical variance adjustment regardless of the correlation between the measures. When the correlation factor is included, all exhibit similar trends, with better performance when $\rho < 0.5$ and worse performance when $\rho > 0.5$. Finally, if $\rho = 0.5$, all six indexes performed similarly.

Discussion

This report analytically evaluates two controversial 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 continuous measure. The second is to determine the best estimates of the standardized mean difference effect size and its variance when the comparisons are derived from a repeated-measures or a between-groups design.

Choice of Metric When Meta-Analyzing Continuous Measures

Although several statistical methods exist to estimate comparisons of groups at one or more points (Tables 1 to 5), none provide unbiased estimations and, before the current report, the circumstances under which they produce the most optimal statistical inferences has been unknown. In the current simulations, the standardized mean difference outperformed the unstandardized version under a broad set of conditions in terms both of bias (Figures 1–5) and of efficiency (Figures 6–10) under the conditions we have described on our methods section. The standardized mean difference performed better when differences in within-study variability are large, when parametric assumptions are poorly met, and when study sample sizes are small. When the underlying assumptions are better met, choice of standardized vs. unstandardized mean difference mattered little in estimates of the weighted mean effect size (Table 7). Table 8 summarizes which equations performed best in the current research in terms of operationalizing effect sizes and their variances 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 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. Naturally, 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 being 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 8. Findings relevant to meta-analytic practice (effect size and variance choice)

Design Type	Inferential Circumstances	Best Performing Equation and Sources	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) Variance (Table 4, No. 15) Hedges (1981) ²⁵	$d_{hb} = c(N - 2) \frac{\bar{Y}_{Post}^E - \bar{Y}_{Post}^C}{S_{Pooled}}$ $\text{var}_{two-g}(d_{b-t}) = \left(\frac{1}{\tilde{n}}\right) \left(\frac{N-2}{N-4}\right) (1 + \tilde{n}d_{b-t}^2) - \frac{d_{b-t}^2}{[c(N-2)]^2},$
	For normally distributed data with equal variances	UMD (Table 5, No. 19); Lipsey & Wilson (2001) ¹⁸	$UMD = \bar{Y}_{Post}^E - \bar{Y}_{Post}^C$
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 et al. (1981) ¹¹	$d_{tra} = t_d \sqrt{\frac{1}{n} 2(1 - r_{Pre,Post})}$
		SMD d_b (Table 1, No. 3); Becker (1988) ¹⁹	$d_b = c(n-1) \frac{\bar{Y}_{Post} - \bar{Y}_{Pre}}{S_{Pre}}$
		Variance (Table 3, No. 13); Becker (1988) ¹⁹	$\text{var}_{one-g}(d_b) = \left(\frac{2(1-r_{pre,post})}{n}\right) \left(\frac{n-1}{n-3}\right) \left(1 + \frac{n}{2(1-r_{pre,post})} d_b^2\right) - \frac{d_b^2}{[c(100-1)]^2}$
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) ³	$d_{tch} = t_d \sqrt{\frac{1}{n}}$
		SMD d_g (Table 1, No. 4); Gibbons et al. (1993) ²³	$d_g = c(n-1) \frac{\bar{Y}_{Diff}}{S_{Diff}}$
		Variance (Table 3, No. 14); Gibbons et al., 1993) ²³	$\text{var}_{one-g}(d_g) = \left(\frac{1}{n}\right) \left(\frac{n-1}{n-3}\right) (1 + nd_g^2) - \frac{d_g^2}{[c(n-1)]^2}$

Table 8. Findings relevant to meta-analytic practice (effect size and variance choice) (continued)

Design Type	Inferential Circumstances	Best Performing Equation and Sources	Equation
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	$d_b = c(N - 2) \left[\frac{\bar{Y}_{Post}^E - \bar{Y}_{Pre}^E}{S_{Pre}^E} - \frac{\bar{Y}_{Post}^C - \bar{Y}_{Pre}^C}{S_{Pre}^C} \right]$
		SMD d_{hw} (Table 2, No. 9); Huedo-Medina & Johnson (2011) ²⁴	$d_{hw} = c(N - 2) \left[\frac{\bar{Y}_{Post}^E - \bar{Y}_{Pre}^E}{S_{within-pool}^E} - \frac{\bar{Y}_{Post}^C - \bar{Y}_{Pre}^C}{S_{within-pool}^C} \right]$
		Variance (Table 4, No. 15); Hedges (1981) ²⁵	$\text{var}_{two-g}(d_{b-t}) = \left(\frac{1}{\tilde{n}} \right) \left(\frac{N-2}{N-4} \right) (1 + \tilde{n}d_{b-t}^2) - \frac{d_{b-t}^2}{[c(N-2)]^2},$
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) ²³	$d_g = c(N - 2) \left[\frac{\bar{Y}_{Diff}^E}{S_{Diff}^E} - \frac{\bar{Y}_{Diff}^C}{S_{Diff}^C} \right]$
		SMDs d_{s2} or d_{s3} (Table 2, Nos. 11 and 12); Shadish et al. (1999) ²⁶	$d_{s2} = \frac{\bar{Y}_{Post}^E - \bar{Y}_{Post}^C}{S_{ANOVA}}$ $d_{s3} = \frac{\bar{Y}_{Post}^E - \bar{Y}_{Post}^C}{S_{ANCOVA}}$
		Variance (Table 4, No. 15); Hedges (1981) ²⁵	$\text{var}_{two-g}(d_{b-t}) = \left(\frac{1}{\tilde{n}} \right) \left(\frac{N-2}{N-4} \right) (1 + \tilde{n}d_{b-t}^2) - \frac{d_{b-t}^2}{[c(N-2)]^2},$

The second solution for clinical interpretation hinges on effect size standards, which are made possible by using the standardized mean difference effect size. Specifically, Cohen^{30,35} 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, p. 156). Thus, if a standardized mean difference exceeds 0.50, then it is likely to be readily noticeable to the careful practitioner. If it is smaller, it is unlikely to be noticeable without the aid of statistics. In other words, if at least a medium amount of improvement has occurred between two observations, it should be noticeable in practice. Similarly, if a trial yielded a medium effect size and one encountered individuals who had been in either the treatment group or the control group, one could notice differences between them. 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. Finally, note that interventions with an average small effect can have very large public health effects if they apply to large part of the population, even if they are not noticeable by clinicians.

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

Tables 1 to 5 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^{11,21,26} or exclude it.^{3,19,23,26} Despite the disagreement about use of the correlation in calculating the ES, all solutions except Gibbons et al.,²³ 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 0.5 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 with the variability of change scores and the effect size in raw-score metric as the mean difference between conditions compared with 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 variability. 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 considers 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 study 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). In general, for one-group repeated-measures designs all ES equations behaved well, but Table 8 lists those that performed the best under certain conditions.

Based on our results, selections of a formula for repeated measures can have considerable 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.

It is also possible that our results favoring standardized mean differences over their unstandardized counterparts were in part determined by the design of our simulations that are more conditioned to the first one than to the latter. A future simulation assuming an unstandardized parametric effect size would be a useful replication and check of this possibility. Our simulation also does not provide direct evidence about the advisability of mixing ESs from between- and within-group designs in the same meta-analysis. Some sources argue against the practice (see Lipsey and Wilson, 2001)¹⁸ and others suggest that it is acceptable (for example, see Morris and DeShon, 2002 and Johnson and Eagly, 2000).^{22,34} Future research should directly address these issues.

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 subsection, 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. The unstandardized difference's relatively 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 variances of the measures 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. Other important aspects 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 may statistically benefit the standardized version more than the unstandardized counterpart.

References

1. Sutton AJ, Duval SJ, Tweedie RL, et al. Empirical assessment of effect of publication bias on meta-analyses. *BMJ*. 2000;320(7249):1574-7. PMID: 10845965.
2. Rothstein H, Sutton AJ, Borenstein M. Publication bias in meta-analysis. In: Rothstein HR, Sutton AJ, & M. Borenstein M, eds. *Publication bias in meta-analysis: prevention, assessment and adjustments*. Chichester, UK: Wiley; 2005.
3. Rosenthal R. *Meta-analytic procedures for social research*. Rev. ed. Newbury Park: Sage Publications; 1991.
4. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ*. 2001;323(7303):42-6. PMID: 11440947.
5. Johnson BT, Boynton MH. Cumulating evidence about the social animal: meta-analysis in social-personality psychology. *Society & Personality Psychology Compass*. 2008;2:817-41.
6. Hunter JE, Schmidt FL. *Methods of meta-analysis: correcting error and bias in research findings*. 2nd ed. Thousand Oaks, CA: Sage; 2004.
7. Huedo-Medina TB, Johnson BT. *Modelos Estadísticos en Meta-análisis [Statistical Models in Meta-analysis]*: La Coruña, Spain: Netbiblio; 2010.
8. Hedges LV, Olkin I. *Statistical methods for meta-analysis*. Orlando: Academic Press; 1985.
9. Cooper HM, Hedges LV, Valentine JC. *The handbook of research synthesis*. New York: Russell Sage Foundation; 2009.
10. Cooper H. *Integrative research: A guide for literature reviews (3rd ed.)*. Newbury Park, CA: Sage; 1998.
11. Glass GV, McGaw B, Smith ML. *Meta-analysis in social research*. Beverly Hills: Sage Publications; 1981.
12. Glaser RR. *Accuracy of ES calculation methods for repeated measures [dissertation]*. University of Memphis; 2002.
13. Ray JW, Shadish WR. How interchangeable are different estimators of effects size? *J Consult Clin Psychol*. 1996;64:1316-25. PMID: 8991318.
14. Bond CF Jr, Wiitala WL, Richard FD. Meta-analysis of raw mean differences. *Psychol Methods*. 2003;8:406-18.
15. Tukey JW. Analyzing data: sanctification or detective work? *Am Psychol*. 1969;24:83-91.
16. Blalock HM. *Causal inferences in non-experimental research*. Chapel Hill: University of North Carolina Press; 1964.
17. Pedhazur EJ. *Multiple Regression in Behavioral Research: Explanation and Prediction*. New York: Holt, Rinehart and Winston; 1997.
18. Lipsey MW, Wilson DB. *Practical Meta-analysis*. Vol 49. Thousand Oaks, CA: Sage Publications; 2001.
19. Becker BJ. Synthesizing standardized mean-change measures. *Br J Math Stat Psychol*. 1988;41:257-278.
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. *Psychol Methods*. 1996;1:170-177.
22. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-group designs. *Psychol Methods*. 2002;7:105-125. PMID: 11928886.
23. Gibbons RD, Hedeker DR, Davis JM. Estimation of ES from a series of experiments involving paired comparisons *J Educ Stat*. 1993;18:271-279.
24. Huedo-Medina T, Johnson BT. Standardized mean difference ES estimations for repeated-measures with continuous measures. 2011.
25. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Behav Stat*. 1981 June 20;6(2):107-128.

26. Shadish WR, Robinson L, Lu C. ES: A computer programs for ES calculation. St Paul, MN: Assesment Systems Corporation; 1999.
27. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans Model Comput Simul.* 1998;8(1):3-30.
28. Wilcox RR. New designs in analysis of variance. *Annu Rev Psychol.* 1987;38:29-60.
29. 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.
30. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
31. Fleishman AI. A method for simulating nonnormal distributions. *Psychometrika.* 1978;43:521-31.
32. Overton RC. A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychol Methods.* 1998;3:354-79.
33. Biggerstaff BJ, Tweedie RL. Incorporating variability estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med.* 1997;16:753- 68. PMID: 9131763.
34. Johnson BT, Eagly AH. Quantitative synthesis of social psychological research. In: Reis HT and Judd CM, eds. *Handbook of Research Methods in Social and Personality Psychology.* London: Cambridge University Press; 2000: 496-528.
35. Cohen J. Quantitative methods in psychology: A power primer. *Psychol Bull.* 1992;112(1):155-9.

Glossary of Terms

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 with the variability of change scores.

Coverage: The proportion of replications for which the 95% confidence interval for each index did not include the null value, $ES = SMD = UMD = 0$.

Effect size (ES): The magnitude or degree of the association between two variables. In the current investigation, comparisons between groups, across time, or both, are used (e.g., standardized mean difference; unstandardized mean difference).

Efficiency: A measure of the optimality of an estimator that reaches the closest value to the parameter with the minimum variance. 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. Bias and efficiency are, in effect, components of the MSE.

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).

Pooled standard deviation: The sample-size weighted mean standard deviations of two or more groups.

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, so using two independent samples (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.

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 reflecting the change in means for two independent groups (repeated-measures, between-groups version) or the comparison of the means at post-test 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.

Following are the Greek terms that appear in this report:

Term	Definition
δ	The parametric difference between two groups
$\hat{\delta}_j$	The sample estimate of population parameter δ for the j_{th} replication
μ_{post}^E	Parametric mean at post-test for the experimental group
μ_{Pre}^E	Parametric mean at pretest for the experimental group
μ_{Post}^C	Parametric mean at post-test for the control group
μ_{Pre}^C	Parametric mean at pretest for the control group
σ_{Pre}^2	Parametric variance at pretest
σ_{Post}^2	Parametric variance at post-test
$\sigma_{Pre,Post}$	Covariance between pre- and post-test
ρ	Parametric correlation
$\rho_{pre-post}$	Parametric correlation between pre-and post-test
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, excluding the correlation factor $2(1 - r)$ in its variance estimation
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 $2(1 - r)$ in its variance estimation.

Following are the Latin abbreviations used in this report (continued):

Term	Definition
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 1, No. 1 for details and elements of the equation)
ES	Effect size
HAM-D	Hamilton rating scale of depression
k	Number of studies
MHAM-D	Mean score on the HAM-D
M_C	Mean for control group
M_E	Mean for experimental group
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
SE _{UMD}	The standard error for the UMD
SE _{SMD}	The standard error for the SMD
SMD	Standardized mean difference (d)
UMD	Unstandardized mean difference (Equation 19, Table 5)
$var_{one-g}(d_b)$	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)
$var_{one-g}(d_g)$	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)
$var_{two-g}(d_b)$	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)
$var_{two-g}(d_{b,t})$	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)
$var_{two-g}(d_g)$	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)
$var_{two-g}(d_{g,t})$	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)
Var _{UMD}	The variance estimate for the UMD
Var _{SMD}	The variance estimate for the SMD
Y^C	Control outcome
Y^E	Experimental outcome

Appendix A. Bias and Efficiency Results for Standardized and Raw Mean Differences (Specific Aim 1)

This appendix contains the bias and efficiency results in relation to the two different effect sizes metrics, unstandardized (UMD) versus standardized (SMD) mean differences, with results divided by simulation conditions that related either to bias, efficiency, or both.

<i>Skewness/Kurtosis</i>	δ	n	<i>SMD</i>	<i>UMD</i>
0/3	0.25	30	-0.0003	0.0210
		50	-0.0005	0.0239
		80	-0.0003	0.0261
	0.50	30	-0.0003	0.0284
		50	-0.0005	0.0309
		80	-0.0003	0.0333
	0.80	30	-0.0003	0.0078
		50	-0.0005	0.0099
		80	-0.0003	0.0122
0.75/0	0.25	30	-0.1021	0.1389
		50	-0.1035	0.1322
		80	-0.1043	0.1346
	0.50	30	-0.1021	0.1605
		50	-0.1035	0.1639
		80	-0.1043	0.1667
	0.80	30	-0.1021	0.1923
		50	-0.1035	0.1955
		80	-0.1043	0.1985
1.75/3.75	0.25	30	0.2001	0.2369
		50	-0.2104	0.2404
		80	-0.2002	0.2431
	0.50	30	0.2001	0.2925
		50	-0.2104	0.2967
		80	-0.2002	0.3001
	0.80	30	0.2001	0.3767
		50	-0.2104	0.3812
		80	-0.2002	0.3848

<i>Skewness/Kurtosis</i>	δ	<i>n</i>	<i>SMD</i>	<i>UMD</i>
		30	0.0001	0.0175
	0.25	50	0.0019	0.0257
		80	0.0017	0.0279
0/3		30	0.0001	0.0259
	0.50	50	0.0019	0.0334
		80	0.0017	0.0353
		30	0.0001	0.0066
	0.80	50	0.0019	0.0131
		80	0.0017	0.0146
0.75/0		30	0.1001	0.1750
	0.25	50	0.1019	0.1839
		80	0.1017	0.1864
		30	0.1001	0.1573
	0.50	50	0.1019	0.1663
		80	0.1017	0.1688
		30	0.1001	0.1907
	0.80	50	0.1019	0.1991
		80	0.1017	0.2012
		30	0.1001	0.1325
	0.25	50	0.1019	0.1421
		80	0.1017	0.1448
1.75/3.75		30	0.1001	0.2887
	0.50	50	0.1019	0.2992
		80	0.1017	0.3023
		30	0.1001	0.3748
	0.80	50	0.1019	0.3851
		80	0.1017	0.3879

Skewness/Kurtosis	δ	n	SMD	UMD
0/3	0.25	30	0.0001	0.0175
		50	0.0019	0.0257
		80	0.0017	0.0279
	0.50	30	0.0001	0.0259
		50	0.0019	0.0334
		80	0.0017	0.0353
	0.80	30	0.0001	0.0066
		50	0.0019	0.0131
		80	0.0017	0.0146
0.75/0	0.25	30	0.1001	0.1750
		50	0.1019	0.1839
		80	0.1017	0.1864
	0.50	30	0.1001	0.1573
		50	0.1019	0.1663
		80	0.1017	0.1688
	0.80	30	0.1001	0.1907
		50	0.1019	0.1991
		80	0.1017	0.2012
1.75/3.75	0.25	30	0.1001	0.1325
		50	0.1019	0.1421
		80	0.1017	0.1448
	0.50	30	0.1001	0.2887
		50	0.1019	0.2992
		80	0.1017	0.3023
	0.80	30	0.1001	0.3748
		50	0.1019	0.3851
		80	0.1017	0.3879

		0	0.0001	0.0175
	1:1	0.08	0.0019	0.0257
		0.32	0.0017	0.0279
0/3		0	0.0001	0.0259
	1:2	0.08	0.0019	0.0334
		0.32	0.0017	0.0353
		0	0.0001	0.0366
	4:1	0.08	0.0019	0.0431
		0.32	0.0017	0.0546
		0	0.1001	0.1750
	1:1	0.08	0.1019	0.1839
		0.32	0.1017	0.1864
0.75/0		0	0.2101	0.3573
	1:2	0.08	0.2119	0.3663
		0.32	0.2117	0.3688
		0	0.3101	0.4907
	4:1	0.08	0.3219	0.4991
		0.32	0.3317	0.4012
		0	0.1001	0.1325
	1:1	0.08	0.1019	0.1421
		0.32	0.1017	0.1448
1.75/3.75		0	0.3101	0.4887
	1:2	0.08	0.3119	0.4992
		0.32	0.3117	0.4023
		0	0.3801	0.5748
	4:1	0.08	0.3819	0.5851
		0.32	0.3917	0.5879

Note. E:C=Variance of experimental group relative to variance of control group.

Skewness/Kurtosis	δ	n	SMD	UMD
0/3	0.25	30	0.1104	0.1108
		50	0.0492	0.0500
		80	0.0276	0.0281
	0.50	30	0.1104	0.1109
		50	0.0492	0.0500
		80	0.0276	0.0282
	0.80	30	0.1104	0.1106
		50	0.0492	0.0499
		80	0.0276	0.0281
0.75/0	0.25	30	0.1868	0.2117
		50	0.0718	0.1504
		80	0.0690	0.1284
	0.50	30	0.1868	0.2131
		50	0.0718	0.1510
		80	0.0690	0.1287
	0.80	30	0.1868	0.2137
		50	0.0718	0.1513
		80	0.0690	0.1289
1.75/3.75	0.25	30	0.2286	0.3126
		50	0.1568	0.2508
		80	0.1318	0.2286
	0.50	30	0.2286	0.3157
		50	0.1568	0.2522
		80	0.1318	0.2294
	0.80	30	0.2286	0.3177
		50	0.1568	0.2531
		80	0.1318	0.2299

Skewness/Kurtosis	δ	n	SMD	UMD
0/3	0.25	30	0.1104	0.1118
		50	0.0492	0.0505
		80	0.0276	0.0284
	0.50	30	0.1104	0.1117
		50	0.0492	0.0504
		80	0.0276	0.0284
	0.80	30	0.1104	0.1112
		50	0.0492	0.0502
		80	0.0276	0.0283
0.75/0	0.25	30	0.1867	0.2127
		50	0.0718	0.1509
		80	0.0690	0.1286
	0.50	30	0.1867	0.2139
		50	0.0718	0.1514
		80	0.0690	0.1289
	0.80	30	0.1867	0.2143
		50	0.0718	0.1516
		80	0.0690	0.1290
1.75/3.75	0.25	30	0.2286	0.3136
		50	0.1568	0.2513
		80	0.1318	0.2289
	0.50	30	0.2286	0.3165
		50	0.1568	0.2526
		80	0.1318	0.2296
	0.80	30	0.2286	0.3183
		50	0.1568	0.2534
		80	0.1318	0.2301

	Skewness/Kurtosis	δ	n	SMD	UMD
0/3	0.25		30	0.0552	0.1149
			50	0.0246	0.0519
			80	0.0138	0.0292
	0.50		30	0.0552	0.1143
			50	0.0246	0.0516
			80	0.0138	0.0290
	0.80		30	0.0552	0.1132
			50	0.0246	0.0511
			80	0.0138	0.0288
0.75/0	0.25		30	0.0983	0.2158
			50	0.0759	0.1523
			80	0.0445	0.1294
	0.50		30	0.0983	0.2165
			50	0.0759	0.1526
			80	0.0445	0.1296
	0.80		30	0.0983	0.2162
			50	0.0759	0.1525
			80	0.0445	0.1295
1.75/3.75	0.25		30	0.1642	0.3167
			50	0.1284	0.2527
			80	0.1159	0.2297
	0.50		30	0.1642	0.3191
			50	0.1284	0.2538
			80	0.1159	0.2303
	0.80		30	0.1642	0.3202
			50	0.1284	0.2543
			80	0.1159	0.2306

Skewness/Kurtosis	Ratio of variances (E:C)	τ^2	SMD	UMD
0/3	1:1	0	0.0001	0.0175
		0.08	0.0019	0.0257
		0.32	0.0017	0.0279
	1:2	0	0.0001	0.0259
		0.08	0.0019	0.0334
		0.32	0.0017	0.0353
	4:1	0	0.0001	0.0066
		0.08	0.0019	0.0131
		0.32	0.0017	0.0146
0.75/0	1:1	0	0.0111	0.0750
		0.08	0.0119	0.0839
		0.32	0.0117	0.0864
	1:2	0	0.0101	0.1573
		0.08	0.0119	0.1663
		0.32	0.0117	0.1688
	4:1	0	0.0101	0.1907
		0.08	0.0119	0.1991
		0.32	0.0117	0.2012
1.75/3.75	1:1	0	0.0180	0.1325
		0.08	0.0216	0.1421
		0.32	0.0225	0.1448
	1:2	0	0.0180	0.2887
		0.08	0.0216	0.2992
		0.32	0.0225	0.3023
	4:1	0	0.0180	0.3748
		0.08	0.0216	0.3851
		0.32	0.0225	0.3879

Note. E:C=Variance of experimental group relative to variance of control group.

Appendix B. Bias, Efficiency, and Theoretical Variance for All Effect Size Indexes and Their Variances (Specific Aim 2)

This appendix provides detailed tables for the results of the Specific Aim 2, including the bias, efficiency, and theoretical variance adjusted for all the effect size indexes and their variances.

<i>n</i>	<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_{tra}</i>	<i>d_{tch}</i>	<i>d_b</i>	<i>d_g</i>
50	.00	0.0001	-0.1491	0.0074	-0.1412	-0.0004	-0.1467
	.25	-0.0007	-0.1497	0.0059	-0.0869	-0.0010	-0.0933
	.50	0.0019	-0.1478	0.0084	0.0084	0.0010	0.0006
	.75	0.0034	-0.1468	0.0084	0.3038	0.0009	0.2914
75	.00	0.0010	-0.1475	0.0066	-0.1418	0.0011	-0.1454
	.25	0.0002	-0.1481	0.0052	-0.0875	-0.0002	-0.0917
	.50	-0.0005	-0.1486	0.0040	0.0040	-0.0009	-0.0011
	.75	0.0017	-0.1471	0.0048	0.2982	0.0001	0.2901
100	.00	0.0012	-0.1469	0.0051	-0.1428	0.0011	-0.1455
	.25	0.0004	-0.1475	0.0034	-0.0890	0.0003	-0.0921
	.50	-0.0003	-0.1480	0.0031	0.0031	-0.0006	-0.0007
	.75	0.0020	-0.1464	0.0043	0.2974	0.0008	0.2913
250	.00	0.0001	-0.1469	0.0014	-0.1455	0.0002	-0.1465
	.25	0.0003	-0.1467	0.0016	-0.0904	0.0004	-0.0916
	.50	-0.0004	-0.1473	0.0010	0.0010	-0.0007	-0.0005
	.75	0.0001	-0.1469	0.0014	0.2928	-0.0005	0.2904
500	.00	0.0002	-0.1466	0.0011	-0.1457	0.0005	-0.1462
	.25	0.0010	-0.1460	0.0018	-0.0903	0.0010	-0.0909
	.50	0.0003	-0.1465	0.0014	0.0014	0.0000	0.0006
	.75	0.0000	-0.1467	0.0005	0.2913	-0.0005	0.2902
750	.00	-0.0001	-0.1467	0.0003	-0.1462	0.0000	-0.1466
	.25	-0.0006	-0.1470	-0.0001	-0.0918	-0.0008	-0.0923
	.50	0.0002	-0.1465	0.0005	0.0005	0.0001	0.0000
	.75	0.0007	-0.1461	0.0008	0.2919	0.0006	0.2911

Note. See Tables 1 and 2 in the main report for definitions of each effect size index. Bias is defined as the *M* difference between the estimate of the ES and the parametric value. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom; *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom; *d_{tch}*=the standardized mean difference in change-score metric from *t_d*; *d_{tra}*=standardized mean difference in raw-score metric form *t_d*.

<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_{tra}</i>	<i>d_{tch}</i>	<i>d_b</i>	<i>d_g</i>
.00	0.0004	-0.1473	0.0037	-0.1439	0.0037	-0.1462
.25	0.0001	-0.1475	0.0030	-0.0893	0.0032	-0.0920
.50	0.0002	-0.1475	0.0031	0.0031	0.0031	-0.0002
.75	0.0013	-0.1467	0.0034	0.2959	0.0035	0.2908

Note. See Tables 1 and 2 in the main report for definitions of each effect size index. Bias is defined as the *M* difference between the estimate of the ES and the parametric value. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom; *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom; *d_{tch}*=the standardized mean difference in change-score metric from *t_d*; *d_{tra}*=standardized mean difference in raw-score metric form *t_d*.

<i>n</i>	<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_{tra}</i>	<i>d_{tch}</i>	<i>d_b</i>	<i>d_g</i>
50	.00	0.0413	0.0206	0.0426	0.0224	0.0432	0.0217
	.25	0.0303	0.0152	0.0312	0.0222	0.0321	0.0215
	.50	0.0212	0.0106	0.0217	0.0230	0.0226	0.0223
	.75	0.0101	0.0051	0.0103	0.0274	0.0107	0.0266
75	.00	0.0269	0.0135	0.0275	0.0144	0.0282	0.0141
	.25	0.0210	0.0105	0.0214	0.0151	0.0221	0.0147
	.50	0.0146	0.0073	0.0149	0.0158	0.0155	0.0155
	.75	0.0066	0.0033	0.0067	0.0176	0.0070	0.0173
100	.00	0.0202	0.0101	0.0206	0.0108	0.0210	0.0106
	.25	0.0159	0.0079	0.0161	0.0112	0.0167	0.0110
	.50	0.0106	0.0053	0.0108	0.0115	0.0113	0.0113
	.75	0.0051	0.0025	0.0051	0.0131	0.0053	0.0129
250	.00	0.0081	0.0040	0.0081	0.0042	0.0083	0.0042
	.25	0.0062	0.0031	0.0062	0.0043	0.0065	0.0043
	.50	0.0043	0.0022	0.0044	0.0046	0.0045	0.0046
	.75	0.0021	0.0010	0.0021	0.0053	0.0021	0.0053
500	.00	0.0041	0.0020	0.0041	0.0021	0.0042	0.0021
	.25	0.0032	0.0016	0.0032	0.0023	0.0034	0.0022
	.50	0.0021	0.0011	0.0021	0.0023	0.0022	0.0023
	.75	0.0010	0.0005	0.0010	0.0026	0.0010	0.0026
750	.00	0.0027	0.0013	0.0027	0.0014	0.0028	0.0014
	.25	0.0021	0.0011	0.0021	0.0015	0.0022	0.0015
	.50	0.0015	0.0007	0.0015	0.0015	0.0015	0.0015
	.75	0.0007	0.0003	0.0007	0.0018	0.0007	0.0018

Note. See Tables 1 and 2 in the main report for definitions of each effect size index. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom; *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom; *d_{tch}*=the standardized mean difference in change-score metric from *t_d*; *d_{tra}*=standardized mean difference in raw-score metric form *t_d*.

r	d_{hb}	d_{hw}	d_{tra}	d_{tch}	d_b	d_g
.00	0.0172	0.0086	0.0176	0.0092	0.0180	0.0090
.25	0.0131	0.0066	0.0134	0.0094	0.0138	0.0092
.50	0.0091	0.0045	0.0092	0.0098	0.0096	0.0096
.75	0.0043	0.0021	0.0043	0.0113	0.0045	0.0111

Note. See Tables 1 and 2 in the main report for definitions of each effect size index. d_b =Becker's standardized mean difference; d_g =Gibbons's standardized mean difference. d_{hb} =Hedges's standardized mean difference using the between-studies degrees of freedom; d_{hw} =Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom; d_{tch} =the standardized mean difference in change-score metric from t_d ; d_{tra} =standardized mean difference in raw-score metric form t_d .

n	r	d_{hb}	d_{hw}	d_b	d_g
50	.00	0.0375	0.0226	0.0018	0.0217
	.25	0.0284	0.0176	0.0023	0.0120
	.50	0.0171	0.0118	0.0013	0.0017
	.75	0.0037	0.0047	0.0005	-0.0102
75	.00	0.0252	0.0149	0.0012	0.0142
	.25	0.0177	0.0110	0.0004	0.0071
	.50	0.0107	0.0073	0.0001	0.0001
	.75	0.0025	0.0031	0.0003	-0.0067
100	.00	0.0188	0.0110	0.0008	0.0105
	.25	0.0132	0.0081	0.0000	0.0053
	.50	0.0083	0.0056	0.0003	0.0003
	.75	0.0018	0.0022	0.0001	-0.0051
250	.00	0.0075	0.0043	0.0002	0.0042
	.25	0.0054	0.0032	0.0001	0.0021
	.50	0.0032	0.0021	0.0000	0.0000
	.75	0.0007	0.0008	0.0000	-0.0022
500	.00	0.0037	0.0021	0.0001	0.0020
	.25	0.0025	0.0015	-0.0001	0.0009
	.50	0.0016	0.0011	0.0000	0.0000
	.75	0.0004	0.0004	0.0000	-0.0011
750	.00	0.0025	0.0014	0.0001	0.0014
	.25	0.0017	0.0010	0.0000	0.0006
	.50	0.0011	0.0007	0.0000	0.0000
	.75	0.0002	0.0003	0.0000	-0.0008

Note. See Tables 1 and 2 in the main report for definitions of each effect size index. d_b =Becker's standardized mean difference; d_g =Gibbons's standardized mean difference. d_{hb} =Hedges's standardized mean difference using the between-studies degrees of freedom; d_{hw} =Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom.

<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_{hw_nonr}</i>	<i>d_b</i>	<i>d_{b_nonr}</i>	<i>d_g</i>
.00	0.0159	0.0094	0.0007	0.0007	-0.0080	0.0003
.25	0.0115	0.0071	0.0028	0.0005	-0.0039	0.0003
.50	0.0070	0.0048	0.0048	0.0003	0.0003	0.0003
.75	0.0016	0.0019	0.0071	0.0002	0.0054	0.0005

Note. d_{b_nonr} =Becker's standardized mean difference without including the correlation factor $2(1-r)$. d_{hw_nonr} =Hedges's standardized mean difference using the within-study degrees of freedom without including the correlation factor $2(1-r)$. See main report, Tables 1 and 2 for definitions of the other ES indexes.

<i>n</i>	<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_b</i>	<i>d_g</i>	<i>d_{s1}</i>	<i>d_{s2}</i>	<i>d_{s3}</i>
50	.00	0.0042	-0.1492	-0.0004	-0.1467	0.0037	0.0038	-0.1191
	.25	0.0041	-0.149	0.0000	-0.0927	0.0033	0.0041	-0.1143
	.50	0.0036	-0.1487	-0.0002	-0.0006	0.0025	0.0045	-0.0977
	.75	0.0041	-0.1473	0.0003	0.2906	0.0029	0.0066	-0.0527
75	.00	0.0022	-0.1485	-0.0003	-0.1464	0.0022	0.0023	-0.1288
	.25	0.0028	-0.1481	-0.0003	-0.0917	0.0022	0.0028	-0.1230
	.50	0.0023	-0.1479	0.0001	-0.0001	0.0018	0.0031	-0.1059
	.75	0.0026	-0.1471	0.0001	0.2902	0.0018	0.0042	-0.0590
100	.00	0.0020	-0.1477	0.0002	-0.1463	0.0021	0.0021	-0.1330
	.25	0.0023	-0.1475	0.0002	-0.0920	0.0019	0.0023	-0.1274
	.50	0.0017	-0.1475	0.0001	0.0000	0.0013	0.0023	-0.1099
	.75	0.0018	-0.1470	0.0000	0.2903	0.0012	0.0031	-0.0621
250	.00	0.0007	-0.1470	0.0001	-0.1466	0.0007	0.0007	-0.1412
	.25	0.0010	-0.1467	0.0005	-0.0916	0.0010	0.0012	-0.1354
	.50	0.0006	-0.1470	-0.0003	-0.0002	0.0003	0.0007	-0.1173
	.75	0.0007	-0.1468	-0.0003	0.2908	0.0003	0.0011	-0.0676
500	.00	0.0000	-0.1468	0.0002	-0.1464	0.0003	0.0003	-0.1441
	.25	0.0004	-0.1466	0.0002	-0.0916	0.0004	0.0005	-0.1381
	.50	0.0005	-0.1467	-0.0002	0.0004	0.0002	0.0004	-0.1194
	.75	0.0005	-0.1466	-0.0003	0.2905	0.0002	0.0005	-0.0693
750	.00	0.0002	-0.1466	0.0001	-0.1464	0.0003	0.0003	-0.1447
	.25	0.0005	-0.1466	-0.0001	-0.0917	0.0003	0.0003	-0.1389
	.50	0.0003	-0.1466	0.0000	-0.0002	0.0002	0.0003	-0.1204
	.75	0.0004	-0.1464	0.0001	0.2904	0.0003	0.0005	-0.0700

Note. See Tables 1 and 2 in the main report for definitions of each effect size index. Bias is defined as the *M* difference between the estimate of the ES and the parametric value. d_b =Becker's standardized mean difference; d_g =Gibbons's standardized mean difference. d_{hb} =Hedges's standardized mean difference using the between-studies degrees of freedom. d_{hw} =Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom. d_{s1} =Standardized mean difference proposed by Shadish (See Table 2, No. 10). d_{s2} =Standardized mean difference proposed by Shadish using the standard deviation from ANOVA results (See Table 2, No. 11). d_{s3} =Standardized mean difference proposed by Shadish using the standard deviation from ANCOVA results (See Table 2, No. 12).

<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_b</i>	<i>d_g</i>	<i>d_{s1}</i>	<i>d_{s2}</i>	<i>d_{s3}</i>
.00	0.0016	-0.1476	0.0000	-0.1465	0.0016	0.0016	-0.1352
.25	0.0019	-0.1474	0.0001	-0.0919	0.0015	0.0019	-0.1295
.50	0.0015	-0.1474	-0.0001	-0.0001	0.0011	0.0019	-0.1118
.75	0.0017	-0.1469	0.0000	0.2905	0.0011	0.0027	-0.0635

Note. See Tables 1 and 2 in the main report for definitions of each ES index. Bias is defined as the *M* difference between the estimate of the ES and the parametric value. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom. *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom. *d_{s1}*=Standardized mean difference proposed by Shadish (See Table 2, No. 10). *d_{s2}*=Standardized mean difference proposed by Shadish using the standard deviation from ANOVA results (See Table 2, No. 11). *d_{s3}*=Standardized mean difference proposed by Shadish using the standard deviation from ANCOVA results (See Table 2, No. 12). *d_{tch}*=the standardized mean difference in change-score metric from *t_d*. *d_{tra}*=standardized mean difference in raw-score metric form *t_d*.

<i>n</i>	<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_b</i>	<i>d_g</i>	<i>d_{s1}</i>	<i>d_{s2}</i>	<i>d_{s3}</i>
50	.00	0.0027	0.0006	0.0027	0.0013	0.0013	0.0013	0.0016
	.25	0.0028	0.0007	0.0027	0.0017	0.0014	0.0014	0.0016
	.50	0.0026	0.0008	0.0027	0.0026	0.0017	0.0017	0.0015
	.75	0.0027	0.0011	0.0027	0.0068	0.0022	0.0022	0.0017
75	.00	0.0018	0.0004	0.0017	0.0009	0.0009	0.0009	0.0010
	.25	0.0018	0.0004	0.0017	0.0011	0.0009	0.0009	0.0010
	.50	0.0017	0.0005	0.0017	0.0017	0.0011	0.0011	0.0009
	.75	0.0018	0.0007	0.0018	0.0044	0.0014	0.0014	0.0010
100	.00	0.0013	0.0003	0.0013	0.0006	0.0006	0.0006	0.0007
	.25	0.0013	0.0003	0.0013	0.0009	0.0007	0.0007	0.0007
	.50	0.0013	0.0004	0.0013	0.0013	0.0008	0.0008	0.0007
	.75	0.0013	0.0005	0.0013	0.0033	0.0010	0.0010	0.0007
250	.00	0.0005	0.0001	0.0005	0.0003	0.0003	0.0003	0.0003
	.25	0.0005	0.0001	0.0005	0.0003	0.0003	0.0003	0.0003
	.50	0.0005	0.0002	0.0005	0.0005	0.0003	0.0003	0.0003
	.75	0.0005	0.0002	0.0005	0.0013	0.0004	0.0004	0.0003
500	.00	0.0003	0.0001	0.0003	0.0001	0.0001	0.0001	0.0001
	.25	0.0002	0.0001	0.0003	0.0002	0.0001	0.0001	0.0001
	.50	0.0003	0.0001	0.0002	0.0003	0.0002	0.0002	0.0001
	.75	0.0002	0.0001	0.0003	0.0006	0.0002	0.0002	0.0001
750	.00	0.0002	0.0000	0.0002	0.0001	0.0001	0.0001	0.0001
	.25	0.0002	0.0000	0.0002	0.0001	0.0001	0.0001	0.0001
	.50	0.0002	0.0001	0.0002	0.0002	0.0001	0.0001	0.0001
	.75	0.0002	0.0001	0.0002	0.0004	0.0001	0.0001	0.0001

Note. See Tables 1 and 2 in the main report for definitions of each ES index. Bias is defined as the *M* difference between the estimate of the ES and the parametric value. *db*=Becker's standardized mean difference; *dg*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom. *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom. *ds1*=Standardized mean difference proposed by Shadish (See Table 2, No. 10). *ds2*=Standardized mean difference proposed by Shadish using the standard deviation from ANOVA results (See Table 2, No. 11). *ds3*=Standardized mean difference proposed by Shadish using the standard deviation from ANCOVA results (See Table 2, No. 12).

<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_b</i>	<i>d_g</i>	<i>d_{s1}</i>	<i>d_{s2}</i>	<i>d_{s3}</i>
.00	0.0011	0.0003	0.0011	0.0006	0.0006	0.0006	0.0006
.25	0.0011	0.0003	0.0011	0.0007	0.0006	0.0006	0.0006
.50	0.0011	0.0004	0.0011	0.0011	0.0007	0.0007	0.0006
.75	0.0011	0.0005	0.0011	0.0028	0.0009	0.0009	0.0007

Note. See Tables 1 and 2 in the main report for definitions of each ES index. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom. *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom. *d_{s1}*=Standardized mean difference proposed by Shadish (See Table 2, No. 10). *d_{s2}*=Standardized mean difference proposed by Shadish using the standard deviation from ANOVA results (See Table 2, No. 11). *d_{s3}*=Standardized mean difference proposed by Shadish using the standard deviation from ANCOVA results (See Table 2, No. 12).

<i>n</i>	<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_{hw_total}</i>	<i>d_b</i>	<i>d_{b_total}</i>	<i>d_g</i>	<i>d_{g_total}</i>	<i>d_{s1}</i>	<i>d_{s3}</i>
50	.00	0.0395	0.2800	0.0408	0.2581	0.0395	0.2367	0.0202	0.0409	0.0400
	.25	0.0394	0.2697	0.0408	0.2592	0.0395	0.2503	0.0271	0.0408	0.0401
	.50	0.0395	0.2603	0.0407	0.2600	0.0394	0.2601	0.0412	0.0405	0.0402
	.75	0.0395	0.2489	0.0404	0.2593	0.0395	0.2104	0.0818	0.0400	0.0402
75	.00	0.0262	0.2637	0.0270	0.2574	0.0262	0.2358	0.0133	0.0271	0.0266
	.25	0.0262	0.2550	0.0270	0.2549	0.0262	0.2468	0.0178	0.0270	0.0266
	.50	0.0262	0.2470	0.0269	0.2536	0.0262	0.2535	0.0269	0.0268	0.0266
	.75	0.0262	0.2414	0.0268	0.2557	0.0262	0.2095	0.0532	0.0265	0.0267
100	.00	0.0196	0.2549	0.0202	0.2556	0.0196	0.2340	0.0099	0.0202	0.0199
	.25	0.0196	0.2482	0.0202	0.2534	0.0196	0.2453	0.0132	0.0202	0.0199
	.50	0.0196	0.2426	0.0201	0.2534	0.0195	0.2532	0.0199	0.0200	0.0199
	.75	0.0196	0.2388	0.0200	0.2549	0.0196	0.2092	0.0394	0.0198	0.0200
250	.00	0.0078	0.2386	0.0080	0.2520	0.0078	0.2304	0.0039	0.0080	0.0079
	.25	0.0078	0.2364	0.0080	0.2516	0.0078	0.2433	0.0052	0.0080	0.0079
	.50	0.0078	0.2339	0.0080	0.2511	0.0078	0.2511	0.0078	0.0080	0.0079
	.75	0.0078	0.2320	0.0080	0.2513	0.0078	0.2072	0.0155	0.0079	0.0079
500	.00	0.0039	0.2337	0.0040	0.2511	0.0039	0.2297	0.0019	0.0040	0.0039
	.25	0.0039	0.2331	0.0040	0.2514	0.0039	0.2431	0.0026	0.0040	0.0039
	.50	0.0039	0.2316	0.0040	0.2510	0.0039	0.2510	0.0039	0.0040	0.0040
	.75	0.0039	0.2302	0.0040	0.2506	0.0039	0.2072	0.0077	0.0039	0.0040
750	.00	0.0026	0.2317	0.0027	0.2504	0.0026	0.2290	0.0013	0.0027	0.0026
	.25	0.0026	0.2305	0.0027	0.2498	0.0026	0.2415	0.0017	0.0027	0.0026
	.50	0.0026	0.2306	0.0027	0.2506	0.0026	0.2506	0.0026	0.0026	0.0026
	.75	0.0026	0.2302	0.0026	0.2510	0.0026	0.2077	0.0051	0.0026	0.0026

Note. See the text in the main report for definitions of each ES index. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom. *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom. *d_{s1}*=Standardized mean difference proposed by Shadish (See Table 2, No. 10). *d_{s3}*=Standardized mean difference proposed by Shadish using the standard deviation from ANCOVA results (See Table 2, No. 12). *d_{ch}*=the standardized mean difference in change-score metric from *t_d*. *d_{tra}*=standardized mean difference in raw-score metric from *t_d*.

<i>r</i>	<i>d_{hb}</i>	<i>d_{hw}</i>	<i>d_{hw_total}</i>	<i>d_b</i>	<i>d_{b_total}</i>	<i>d_g</i>	<i>d_{g_total}</i>	<i>d_{s1}</i>	<i>d_{s3}</i>
.00	0.0166	0.2504	0.0171	0.2541	0.0166	0.2326	0.0169	0.0172	0.0168
.25	0.0166	0.2455	0.0171	0.2534	0.0166	0.2451	0.0168	0.0171	0.0168
.50	0.0166	0.2410	0.0171	0.2533	0.0166	0.2533	0.0166	0.0170	0.0169
.75	0.0166	0.2369	0.0170	0.2538	0.0166	0.2085	0.0155	0.0168	0.0169

Note. See the text in the main report for definitions of each ES index. Bias is defined as the *M* difference between the estimate of the ES and the parametric value. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hb}*=Hedges's standardized mean difference using the between-studies degrees of freedom. *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom. *d_{s1}*=Standardized mean difference proposed by Shadish (See Table 2, No. 10). *d_{s3}*=Standardized mean difference proposed by Shadish using the standard deviation from ANCOVA results (See Table 2, No. 12). *d_{ch}*=the standardized mean difference in change-score metric from *t_d*. *d_{tra}*=standardized mean difference in raw-score metric form *t_d*.

<i>n</i>	<i>r</i>	<i>d_{hw_total_r}</i>	<i>d_{hw_total_nonr}</i>	<i>d_{b_total_r}</i>	<i>d_{b_total_nonr}</i>	<i>d_{g_total_r}</i>	<i>d_{g_total_nonr}</i>
50	.00	0.0208	0.0408	0.0195	0.0395	0.0202	0.0402
	.25	0.0279	0.0408	0.0266	0.0395	0.0271	0.0400
	.50	0.0423	0.0407	0.0411	0.0394	0.0412	0.0396
	.75	0.0855	0.0404	0.0845	0.0395	0.0818	0.0368
75	.00	0.0137	0.0270	0.0129	0.0262	0.0133	0.0266
	.25	0.0184	0.0270	0.0176	0.0262	0.0178	0.0265
	.50	0.0277	0.0269	0.0269	0.0262	0.0269	0.0262
	.75	0.0556	0.0268	0.0550	0.0262	0.0532	0.0244
100	.00	0.0102	0.0202	0.0096	0.0196	0.0099	0.0199
	.25	0.0136	0.0202	0.0130	0.0196	0.0132	0.0198
	.50	0.0205	0.0201	0.0199	0.0195	0.0199	0.0195
	.75	0.0412	0.0200	0.0408	0.0196	0.0394	0.0182
250	.00	0.0040	0.0080	0.0038	0.0078	0.0039	0.0079
	.25	0.0054	0.0080	0.0051	0.0078	0.0052	0.0079
	.50	0.0081	0.0080	0.0078	0.0078	0.0078	0.0078
	.75	0.0162	0.0080	0.0160	0.0078	0.0155	0.0073
500	.00	0.0020	0.0040	0.0019	0.0039	0.0019	0.0039
	.25	0.0027	0.0040	0.0025	0.0039	0.0026	0.0039
	.50	0.0040	0.0040	0.0039	0.0039	0.0039	0.0039
	.75	0.0080	0.0040	0.0079	0.0039	0.0077	0.0036
750	.00	0.0013	0.0027	0.0013	0.0026	0.0013	0.0026
	.25	0.0018	0.0027	0.0017	0.0026	0.0017	0.0026
	.50	0.0027	0.0027	0.0026	0.0026	0.0026	0.0026
	.75	0.0053	0.0026	0.0053	0.0026	0.0051	0.0024

Note. See the text in the main report for definitions of each ES index. *d_b*=Becker's standardized mean difference; *d_g*=Gibbons's standardized mean difference. *d_{hw}*=Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom.

<i>r</i>	$d_{hw_total_r}$	$d_{hw_total_nonr}$	$d_{b_total_r}$	$d_{b_total_nonr}$	$d_{g_total_r}$	$d_{g_total_nonr}$
.00	0.0087	0.0171	0.0082	0.0166	0.0084	0.0169
.25	0.0116	0.0171	0.0111	0.0166	0.0113	0.0168
.50	0.0176	0.0171	0.0170	0.0166	0.0171	0.0166
.75	0.0353	0.0170	0.0349	0.0166	0.0338	0.0155

Note. See the text in the main report for definitions of each ES index. Bias is defined as the *M* difference between the estimate of the ES and the parametric value. d_b =Becker's standardized mean difference; d_g =Gibbons's standardized mean difference. d_{hw} =Huedo-Medina and Johnson's standardized mean difference using the within-study degrees of freedom.