

Campbell Collaboration Statistical Analysis Policy Brief

Prepared for the Campbell Collaboration Steering Committee by:

Betsy Jane Becker

Michigan State University

Larry V. Hedges

The University of Chicago

Therese D. Pigott

Loyola University of Chicago

Campbell Collaboration Statistical Analysis Policy Brief

Executive Summary:

Systematic reviews of the effects of interventions and relations among variables often rely on statistical summaries of the results of primary studies. Because Campbell Collaboration (C2) systematic reviewers are likely to face a variety of statistical issues in conducting reviews, this policy brief attempts to:

1. identify the key issues that are confronted by C2 systematic reviewers who want to synthesize the results of studies statistically,
2. outline possible ways that statistical procedures might be used, and
3. provide exemplars of how these methods might be used.

In this brief we address six key issues concerning statistical analysis, and make proposals for C2 policies for each. A summary of the issues and our proposals follows.

1. When conducting a research synthesis, is it ever appropriate for a C2 reviewer to do a review *without* statistically integrating the results of studies? If yes, what are the characteristics of the literature that make this permissible?

Proposal: Study findings should be represented as effect sizes (i.e., indices of treatment impact or relationship strength) in C2 reviews whenever the studies being summarized present quantitative findings. Statistical integration should only be used in any C2 review (or any part of a C2 review) where a summary conclusion from at least two studies is desired, the studies and effect sizes are sufficiently similar to justify integration, and the number of studies is sufficient to support the analysis used in that statistical integration.

2. When statistical integration is used in a C2 review, are there certain statistical

procedures that should routinely be carried out? If so, what are they?

Proposal: Statistical summaries of average effects and variation in effects should be computed (and reported) for fixed-effects, random-effects or both types of analyses. The specific statistics used will depend on whether the review is aimed at (a) estimating a mean effect across studies, (b) examining the variation in effect-size estimates across studies, or (c) fitting a model of effect-size variation.

3. When systematic reviews retrieve and code characteristics of statistical analyses, what characteristics of the analyses should routinely be coded, and, if possible, examined for their impact on the outcomes of studies?

Proposal: Reviewers should code (a) characteristics of the statistical analyses used in the primary study and (b) details about the computations used for the effect size derived from that study. C2 takes the position that it is important to document specific statistical procedures and methods for computing an effect size just as it is important to code study design differences. Coding of statistical procedures allows the use of sensitivity analyses as a method for examining how differences in statistical methods of studies or effect-size computations influence the results of the systematic review.

4. Should multiple (nonindependent) effect-size estimates from the same study ever be used in a C2 synthesis?

Proposal: Reviewers should not ignore dependence among study outcomes. They should use *some* procedure to deal with dependence, describing and giving a justification for that procedure, even if it is ad hoc. Simple approaches such as dropping or combining outcomes or using sensitivity analyses may make sense if the amount of dependent data is small. More sophisticated analyses may be needed if multivariate data are prevalent in the review. In such cases the reviewer must assess the similarity of studies and availability of reliable information on the extent of dependence.

5. Should C2 have a role in advancing cross-design synthesis methods (e.g., propensity scoring and alternatives)? What must be considered if/when reviewers combine estimates of effect from randomized trials with estimates of effect based on other designs, such as surveillance systems, passive observational studies, etc?

Proposal: In some syntheses results from subsets of studies in the synthesis will not be comparable. In such cases reviewers should not summarize across the designs, but rather should report both sets of results separately. In other cases where effects are more comparable, the reviewer may wish to summarize across designs as well as provide separate results by design. Assumptions underlying such comparisons should be made explicit, and the reviewer should critically examine the data for the possibility of design-related differences in effects. Further, when such comparisons are made, the type of design should be tested as a moderator variable and separate results should be reported.

Furthermore, while the primary focus of C2 is on matters directly related to

research cumulation, the study and careful application of methods of cross-design synthesis is consistent with the goals of C2.

6. What should be the role of C2's Social, Psychological, Educational and Criminological Trials Register (SPECTR) in supporting or informing the statistical research that might be done in the Campbell context?

Proposal: The Steering Committee should endorse the use of SPECTR for research on normative methodological and reporting practice in relevant research domains, improving information for imputation in effect-size computation, and studying associations between synthesis methods and results.

Campbell Collaboration

Statistical Analysis Policy Brief

Preamble

The purpose of this Policy Brief is to address issues concerning statistical analysis that will arise in systematic reviews submitted as Campbell Collaboration (C2) reviews.

Additionally, this document does not give an exhaustive treatment of all the statistical issues that may arise in the context of systematic reviewing, nor does it give detailed steps in the analysis process. Clearly a document that attempted to be exhaustive would in no way be "brief." Consequently we intend this to be an evolving document, with periodic addenda anticipated as experience with C2 systematic reviews accumulates.

Procedures

The Statistical Analysis Policy Brief was initiated by the Campbell Collaboration Methods Group, which assigned the task of developing the Brief to Betsy Becker, Larry Hedges and Therese Pigott. The team of authors developed an initial draft in January 2002. This draft was circulated to two initial outside experts (Michael Borenstein and Mark Lipsey) and input was sought on the key questions for the brief at the C2 meetings in Philadelphia (February 21-22, 2002). Modifications were made on the basis of the feedback from the two initial reviewers and the input from the C2 meeting. Then a second draft was circulated to three additional readers (Harris Cooper, Kyle Fahrback and Ingram Olkin). Feedback from Cooper, Fahrback and Olkin was incorporated into this document. The result of the process to date is the present document, which is presented for further criticism and comment by a wider audience.

Introduction

Systematic reviews of the effects of interventions and relations among variables often rely on statistical summaries of the results of primary studies. Because C2 systematic reviewers are likely to face a variety of statistical issues in conducting reviews, this policy brief has attempted to:

1. identify the key issues that are confronted by C2 systematic reviewers who want to synthesize the results of studies statistically,
2. outline possible ways that statistical procedures might be used, and
3. provide exemplars of how these methods might be used.

We should note that a variety of guidelines for the conduct of reviews already exist. Some have been suggested by scholars of the review process (e.g., the checklist in Light & Pillemer, 1984) while others were developed and approved by panels of experts, primarily in the medical arena (e.g., Cook, Sackett & Spitzer, 1995; Moher et al., 1999; Stroup et al., 2000). In addition more detailed guidance concerning the data analysis process is available in such documents as the Cochrane Collaboration Reviewers' Handbook (<http://www.cochrane.org/cochrane/hbook.htm>), though eventually a similar handbook may be developed for Campbell Collaboration reviews. In sum, this brief does not aim to cover all of the issues dealt with in these comprehensive sets of guidelines, nor can this brief cover all of the statistical issues arising in meta-analysis. Interested readers should consult one of the many books available on the process of meta-analysis, a list of which is included in an Appendix A.

Key Issues

The panel, in collaboration with the C2 Steering Committee, identified the following six key issues for consideration in a C2 Policy Brief on Statistical Analysis:

1. When conducting a research synthesis, is it ever appropriate for a C2 reviewer to do a review *without* statistically integrating the results of studies? If yes, what are the characteristics of the literature that make this permissible?
2. When statistical integration is used in a C2 review, are there certain statistical procedures that should routinely be carried out? If so, what are they?
3. When systematic reviews retrieve and code characteristics of statistical analyses, what characteristics of the analyses should routinely be coded, and, if possible, examined for their impact on the outcomes of studies?
4. Should multiple (nonindependent) effect-size estimates from the same study ever be used in a C2 synthesis?

5. Should C2 have a role in advancing cross-design synthesis methods (e.g., propensity scoring and alternatives)? What must be considered if/when reviewers combine estimates of effect from randomized trials with estimates of effect based on other designs, such as surveillance systems, passive observational studies, etc?
6. What should be the role of C2's Social, Psychological, Educational and Criminological Trials Register (SPECTR) in supporting or informing the statistical research that might be done in the Campbell context?

1. When conducting a research synthesis, is it ever appropriate for a C2 reviewer to do a review *without* statistically integrating the results of studies? If yes, what are the characteristics of the literature that make this permissible?

Whether statistical integration is necessary in a synthesis depends on the purpose of that synthesis (see Cooper, 1988 for a discussion of types of syntheses and their purposes). Statistics should be used when the review presents a summary conclusion from at least two studies and a conclusion is desired about effects of one variable on another (that is, for what Cooper calls *integrative* syntheses). Often a summary conclusion will be a statement about the effect of a treatment on an outcome variable, but it could also be a statement about the relation of a risk factor to an outcome.

Prior to *integration* of results of studies (and regardless of the method of integration used), it is necessary to *represent* the results of studies in some form. If the studies selected for review present quantitative findings on effects, the review should represent the findings of each study using an appropriate effect size. The representation of study results in terms of effect sizes makes explicit the study results used in analyses and facilitates comparison of results across studies and across domains. Also if possible the actual effect sizes (or relevant sufficient statistics) should be reported in the review.

If the purpose of the synthesis is to summarize evidence across studies, the studies summarized must (a) examine the same conceptual question at the relevant level of abstraction and (b) be sufficiently similar for their evidence to be combined. If studies are represented by effect sizes and they are sufficiently similar to be combined, statistical methods of integration should be used. If the studies (a) examine the same question at different levels of abstraction, (b) examine conceptually different questions, or (c) are too dissimilar in other ways to make summarizing evidence across studies desirable, then statistical integration across all studies may not be appropriate. In such cases it may, however, be sensible to summarize subsets of studies that could all be included in the same review report.

Summary conclusions may take several forms. Most commonly they are conclusions about the

average or typical effect, or treatment impact. They may also be conclusions about the existence of effects in at least some studies (as in syntheses that screen for unanticipated outcomes or side effects of treatments—where the idea is to determine if there is any evidence of any adverse reactions or unanticipated outcomes). Summary conclusions may also be about variability or consistency of effects (as in syntheses where the question of consistency of treatment effects across implementations is important), which variation of treatment works best, in which setting a particular treatment works best, or for what group or population a treatment works best.

The form of the summary conclusion will determine the appropriate type of statistical integration. Conclusions about the average effect will typically involve estimating a mean effect size of some kind. Conclusions about existence of effects in at least some studies will typically involve tests of the statistical significance of combined results (see Becker, 1987). Conclusions about variability will typically involve analyses of effect-size variance using heterogeneity statistics or variance components. Conclusions about what variation of treatment works best will typically involve analysis of variance (ANOVA) or regression analogues for effect sizes. Note that more complex analyses require larger numbers of studies (for example, moderator analyses involving multiple regression analogues require a larger number of studies than computing a weighted mean). One aspect of appropriateness is whether the number of studies in the review can support the complexity of the analysis.

If the synthesis does not present a summary conclusion of the effects of one variable on another, then statistical integration may be unnecessary. For example, syntheses whose principal purpose is to examine methods or theoretical perspectives used in studies may not need statistics. Similarly, statistical integration may not be needed in syntheses that focus on the range of applications of a treatment. Syntheses that focus on populations or settings (e.g., that look for gaps in who has been studied) may similarly not need statistical summaries. Finally, statistical integration may not be needed in studies that focus on identifying structure in the research literature (such as networks of shared theoretical or methodological perspectives, influence, or the “genealogy” of a research field).

Some syntheses have multiple purposes. Some purposes may require statistical integration (such as summarizing the differences in effects of a treatment found on different types of measures) while others may not (such as presenting the theoretical explanations of why an effect is found on one type of measure and not on another). In syntheses with multiple purposes, statistical integration is necessary in the parts of the synthesis that are integrative.

Proposed Policy: Study findings should be represented as effect sizes (i.e., indices of treatment impact or relationship strength) in C2 reviews whenever the studies under review present quantitative findings. Statistical integration should only be used in any C2 review

(or any part of a C2 review) where (a) a summary conclusion from at least two studies is desired, (b) the studies and effect sizes are sufficiently similar to justify integration, and (c) the number of studies is sufficient to support the analysis used in that statistical integration.

2. When statistical integration is used in a C2 review, are there certain statistical procedures that should routinely be carried out? If so, what are they?

As argued in connection with the previous question, summary conclusions in a systematic review may take three general forms: (a) conclusions about the average or typical effect of the program, (b) conclusions about the variability or consistency of the effects in different studies of the intervention, and (c) conclusions about how the effect of the intervention may vary across settings, research subjects, fidelity of implementation, and other characteristics of studies. Associated with each summary conclusion is a set of statistical procedures that should routinely be carried out. The sections below describe these statistical techniques, and introduce other methods that may prove useful in some C2 reviews. Our recommendations represent only a minimum standard and should not preclude the use of more sophisticated or complex analyses (e.g., Bayesian methods).

Conclusions about the average effect. Many reviews of educational and social interventions are concerned with estimating the size of the intervention effect in a set of studies. Other syntheses may seek information on the typical strength of a relationship or the average incidence of some outcome (e.g., the average dropout rate or other proportion). Most studies will not have identical sample sizes. Therefore, each study's estimate of the effect magnitude varies in its precision due to the variation in the sample size of the study. Given appropriate sampling procedures, studies with larger sample sizes provide more precise estimates of effects than studies with small sample sizes. The C2 Statistics group recommends that C2 reviewers report the weighted mean effect magnitude, weighted by the inverse of the variance of the effect estimate to account for these differences in precision.

As discussed elsewhere (Hedges & Olkin, 1985; Rosenthal, 1994; Haddock, Rindskopf & Shadish, 1998; Lipsey & Wilson, 2001), the variance of effect magnitude estimates depends on a function of the sample size in the study no matter what statistic is being integrated. The C2 Statistics group recommends that, in addition to computing the weighted mean effect, the C2 reviewer provide a 95% confidence interval for the weighted mean effect (where the standard error of the weighted mean effect size is the inverse of the sum of the weights as discussed in Hedges & Olkin, 1985). The confidence interval also allows the reviewer to check whether the weighted mean effect size is statistically different from zero. In order to examine the influence of outliers in the computation of the mean effect size, the C2 Statistics group urges the computation of influence statistics (e.g., as discussed by Greenhouse & Iyengar, 1994).

In some research syntheses, reviewers may hypothesize that the variation among study effect sizes is the result of more than sampling differences between studies. In that case, reviewers may choose to fit random-effects models to the data. For random-effects models, the C2 Statistics group recommends the reporting of the random-effects weighted mean effect size and its associated 95% confidence interval. More will be said about random-effects in the section on advanced topics below.

Conclusions about the variability or consistency of an effect. Though reviewers can simply estimate an average effect for a set of studies, reviewers often want to know how much variation exists in effect estimates across studies. A first step in examining variation in effect-size estimates from studies is a confidence interval plot (sometimes called a forest plot), a common display used in Cochrane Collaboration reviews and described by Light, Singer and Willett (1994). These plots show the estimated effect size from each study along with its 95% confidence interval. The confidence intervals from all studies are lined up on the same plot. The C2 Statistics group recommends presenting confidence interval plots whether or not statistical integration or averaging across studies is desired. Along with the confidence interval plot, the C2 Statistics group recommends the computation of the homogeneity statistic, its associated degrees of freedom and its significance test. The homogeneity statistic can be used to test whether the variation in estimated effect sizes is more than would be expected if all studies were estimating a similar mean effect.

The C2 Statistics group recommends that when fitting a random-effects model, reviewers report the estimate of the variance component, and the test that the variance component is different from zero (including degrees of freedom and significance value), along with the random-effects weighted effect estimates, their associated standard errors, and the 95% random-effects confidence intervals for the effects in a confidence interval plot.

Conclusions about relationships between effect size and study characteristics. Often reviewers wish to know not only whether an effect varies, but also whether that variation is related to characteristics of a study. For example, are particular instantiations of an intervention more effective than others? Is an intervention more effective with certain groups of subjects and/or in certain settings? These questions require the testing of models for moderator variables that might be related to effect magnitude. Two major classes of models for systematic variation among effects are commonly used. The first are termed categorical models of effects, and they are analogous to ANOVA models. Categorical models of effect size are used most often when the reviewer is interested in how effects differ based on discrete groups of studies, such as whether effect estimates differ in studies that use as subjects only women, only men, or a mixed-gender group. The second class includes linear models of effect size, analogous to regression models. These models are used when reviewers are interested in how (a) continuous characteristics of studies or (b) a set of predictors relates to variation in effect-size estimates. For example, a reviewer might wish

to examine how effect sizes differ as a function of the publication date of the primary study, the average age of the subjects, or the duration of a treatment.

For categorical models of effect size, the C2 Statistics group recommends the computation of the weighted mean effect and its associated 95% confidence interval in each discrete group of studies as defined by the categorical factor. As discussed earlier, the weighted mean effect and its confidence interval allow inferences about whether the mean effect within a given group is different from zero. In addition, the C2 Statistics group also recommends the reporting of the homogeneity statistic, its degrees of freedom and significance level within each group. Each within-group homogeneity statistic provides evidence about whether the effect magnitudes within the given group vary more than would be expected if the effects all estimated the same mean. The C2 Statistics group also recommends the computation of the overall between-groups homogeneity statistic which tests whether the groups of studies estimate a single mean versus the alternative hypothesis that at least one group of studies has a mean effect that is different from the other groups. The between-groups homogeneity test is analogous to the between-groups F-test in ANOVA. If random-effects models are used, the C2 Statistics group recommends the computation of the estimate of the variance component within each group, and the test of the variance component along with its degrees of freedom and significance level. In addition, the C2 Statistics group recommends reporting the random-effects weighted mean effect and its 95% confidence interval within each group.

For linear models of effect size, the C2 Statistics group recommends the use of weighted least squares estimates of the association of moderator variables and effect magnitude, with the inverse of the variance of a study's effect as the weighting variable (c.f., Hedges & Olkin, 1985). C2 reviewers should report the weighted regression coefficients and their standard errors, as well as confidence intervals for these coefficients. C2 reviewers should note that while standard statistical packages can produce results of weighted least squares analysis, the printed estimates of the standard errors for the regression coefficients will need to be adjusted as described by Hedges and Olkin (1985). The C2 Statistics group recommends reporting the weighted residual sum of squares statistic, its associated degrees of freedom and significance as a test of model specification. In addition, the C2 Statistics group recommends the examination of the standardized residuals for the weighted regression analysis. When using random-effects models, the C2 Statistics group recommends producing the same statistics as in the fixed-effects linear model, using random-effects weights.

Advanced topic: Fixed- versus random-effects models. Much has been written about the types of generalizations (what Hedges and Vevea (1998) call the inference model) possible from a meta-analysis. The choice of an inference model will determine whether a reviewer will use a fixed-effect analysis or a random-effects analysis. The C2 Statistics group anticipates that most C2 reviewers will, at least initially, base their primary inferences on a fixed-effects analysis. In a fixed-effects analysis, variation between studies in estimates of

effect size are assumed due to sampling error and differences in the studies' methods or characteristics. The inferences supported by fixed-effects analysis concern the sample of studies gathered for the review. Since C2 reviews are intended as exhaustive searches of the literature, the fixed-effects analysis will apply to many C2 reviews. Also, the fixed-effects analysis is the basis for the overall test of homogeneity, and thus serves as a starting point for many analyses.

Primary inferences may be based on random-effects analysis when reviewers believe that the effect sizes estimated from any given study are sampled from an underlying distribution, so that study effects would differ even if the studies had similar characteristics. When a C2 reviewer uses random-effects models, the C2 Statistics group recommends the computation of the random-effects statistics that are analogous to those for fixed-effects models.

Proposed policy: Statistical summaries of average effects and variation in effects should be computed (and reported) for either fixed-effects, random-effects, or both types of analyses. The specific statistics used will depend on whether the review is aimed at (a) estimating a mean effect across studies, (b) examining the variation in effect-size estimates across studies, or (c) fitting a model of effect-size variation. A list of the statistical analyses that should routinely be carried out is given in Appendix B.

3. When systematic reviewers retrieve and code characteristics of statistical analyses, what characteristics of the analyses should routinely be coded, and, if possible, examined for their impact on the outcomes of studies?

Two potential areas of coding concern (a) the analyses conducted in the study itself, and (b) the characteristics of the effect size computed from the study. As Lipsey and Wilson (2001) describe, a reviewer should code the type of statistical procedures used, the significance level, and the direction of the effect. This information provides a check on the value of the effect size if it is needed later in the systematic review process. In some studies, the comparison of interest to the reviewer is not the primary focus of the study. For example, a study may employ a multifactor ANOVA where the comparison of interest (a gender difference, for example) is used as a control or moderator variable. Differences between studies in the statistical procedures used typically result from differences in the research designs employed in the studies. These design differences, as discussed in Question 2, can lead to different assumptions about the nature of the effect-size indices that are derived from the study and possible differences in the magnitudes of computed effect sizes. Sensitivity analyses can shed light on the relation between designs and methods of effect-size computation and the magnitude of the corresponding effect sizes (see, e.g., Greenhouse & Iyengar, 1994).

The second set of codes should describe the effect size computed from each study. A number of sources describe the calculation of effect sizes from various information given in a study (e.g., Haddock, Rindskopf & Shadish, 1998; Hedges & Olkin, 1985; Lipsey & Wilson, 2001; Rosenthal, 1994). How an effect size is computed may affect the analysis and the outcomes of the review (see, e.g, Morris & DeShon, 2002). In some studies, a reviewer may compute a posttest effect size with adjustments for pretest differences whereas in other cases only the unadjusted posttest effect size can be derived. Codes for the type of effect size computed are needed to examine whether effect-size values vary as a function of the computation procedure (and indirectly the research design) employed in the study. The Appendix C gives a list of information about the statistical analysis used in the study, and the type of effect size computed that should be coded.

In addition to these two areas of codes it is important to point out that numerous other characteristics of studies related to statistical issues need to be routinely coded by reviewers. These characteristics will include some that are closely related to statistical issues, for example the cut-points chosen by researchers when they convert continuous variables into dichotomous ones, and with implications for statistical outcomes, for example, the sex, age, and geographic and geopolitical make-up of the treatment and control groups, and the description of the control group.

Proposed policy: Reviewers should code both characteristics of the statistical analyses used in the primary study and details about the computations used for the effect size derived from that study. C2 takes the position that it is important to document specific statistical procedures and methods for computing an effect size, just as it is important to code study design differences. Coding of statistical procedures allows the use of sensitivity analyses to examine whether differences in statistical methods of studies or effect-size computations influence the results of the systematic review.

4. Should multiple (nonindependent) effect-size estimates from the same study ever be used in a C2 synthesis?

The issue underlying this question is that many commonly used procedures for quantitative synthesis assume that the data points to be summarized or analyzed are statistically independent. However, as primary research becomes increasingly complex and multivariate, meta-analysts will encounter multivariate data and dependence as they attempt to synthesize modern primary research.

Explicit dependence among outcomes in meta-analysis can arise in a variety of ways. Most commonly it appears when multiple outcome variables are reported in primary research studies. For example, most studies of the effects of coaching on the Scholastic Aptitude Test (SAT) report results for verbal and quantitative subtests (e.g., Becker, 1990; DerSimonian & Laird, 1983). Typically, however, different studies report different numbers and kinds of outcomes, so across studies the structure of the multiple outcomes is not the same. Dependence may also occur when primary research studies present comparisons of multiple treatment groups to a common control group, even if only a single outcome is observed (Gleser & Olkin, 1994). In such cases dependence results from the fact that the control group mean appears in all contrasts that would be computed.

More subtle sources of dependence may arise when a single primary research source (e.g., one article) reports on multiple (independent) samples that were examined using similar measures, treatments, and the like, or when one researcher and his or her collaborators produce a series of separately reported studies using similar methods and samples. DerSimonian and Laird (1983) examined such effects in the SAT coaching literature.

In primary research the consequences of explicit dependence are well understood and methods for dealing with dependence are easily available (e.g., repeated measures analysis of variance, multivariate techniques). In the realm of research synthesis, however, while some sophisticated approaches are available for dealing with some aspects of dependence (see Becker, 2000, for a review of multivariate meta-analysis approaches), the consequences of accounting for (modeling) dependence or ignoring it are not well understood.

Many researchers have discussed multivariate approaches to meta-analysis (Becker, 1992, 1995; Dunlap et al., 1996; Hedges & Olkin, 1985, chapter 10; Kalaian & Raudenbush, 1996; Raudenbush, Becker & Kalaian, 1988; Timm, 1999). However, if one commits to modeling dependence in meta-analytic data, the information required for most of the available approaches to multivariate meta-analysis (such as matrices of intercorrelations) are often not reported in primary studies. Thus the application of these multivariate approaches may require data imputation, which can be both complex and problematic.

An easily implemented strategy for reducing dependency is to create independent subsets of data for analysis (e.g., Greenwald, Hedges, & Laine, 1996). Separate analyses can then be done for each outcome construct or time-point, and each will be based on independent data points. However, comparisons of results across the data subsets (e.g., comparisons of constructs such as math and verbal outcomes) are still influenced by the dependence that resides in the original data, so this approach is best used only when such comparisons are not of interest, or if the separate constructs have low intercorrelations within studies.

Proposed Policy: The situations in which multivariate data may arise are many and varied, but a wide range of options is available to a reviewer with multivariate meta-analytic data. Reviewers should not ignore dependence among study outcomes. They should use *some* procedure to deal with dependence and describe the procedure clearly, along with their justification for that procedure, even if it is ad hoc.

Simple approaches such as dropping or combining outcomes or using sensitivity analyses (e.g., Greenhouse & Iyengar, 1994) to evaluate the impact of the dependence on results may make sense if the amount of dependent data is small. More sophisticated analyses may be called for if multivariate data are a major part of the evidence in the review. In such cases the reviewer should consider how similar the sets of outcomes are across studies and whether reliable information on the extent of dependence (e.g., intercorrelations among outcome measures) is available.

5. Should C2 have a role in advancing cross-design synthesis methods (e.g., propensity scoring and alternatives)? What must be considered if/when reviewers combine estimates of effect from randomized trials with estimates of effect based on other designs, such as surveillance systems, passive observational studies, etc?

Cross-design synthesis is a term coined by researchers at the US General Accounting Office (GAO) Program Evaluation and Methodology Division (1992) while examining the literature on medical effectiveness. Their initial intent was to draw on the strengths of a variety of data sources in their review by combining data from randomized controlled trials (RCTs) with information from other study designs, particularly "data base" analyses (e.g., hospital patient record data sets).

Boruch and Terhanian (1998) have advanced the idea of cross-design synthesis as a tool for the improvement of study design. They argue that the ideas of cross-design synthesis can help to bridge the gap between experimental studies on interventions and related large-scale data collection efforts in the social sciences (akin to the RCTs and hospital data bases of concern to the GAO). They also give three examples of literatures where cross-design synthesis could be used to understand educational productivity, and examine the issues relevant to each, giving special attention to the populations and treatments studied, the outcomes assessed, and the use of propensity scores in analyses. Thus two purposes are associated with cross-design synthesis - the summary of diverse designs and the improvement of future study designs.

The original form of cross-design synthesis considered by the GAO brings randomized controlled designs and broad population surveys together in one synthesis. A form of cross-design synthesis with less dramatic between-design differences occurs when reviewers gather both randomized and non-randomized studies of a particular intervention, or treatment-control studies and studies examining pretest to posttest change for a treatment sample without the benefit of a comparison or control group.

Two key issues underlie the scenarios outlined above. One is that the data presented in these different designs may lead to the computation of different indices of study effects. In some cases, parallel statistical indices may be available, but in others it may not be possible to compute similar statistical indices (e.g., where some studies use a control group and others do not).

Even if indices for all studies can be computed based on a single formula, differences in study design may require or warrant very different assumptions about the nature and interpretation of the effect indices. For instance, one can compute Glass's classical treatment-control standardized mean difference using posttest means and standard deviations from both randomized and non-randomized treatment-control studies of some intervention, but it is often risky to assume that the groups from the nonrandomized studies were comparable a priori, as researchers typically do when randomization is employed (see Morris & DeShon, 2002, for more discussion of this issue).

With both issues the concern is whether the effect-size values associated with the different design types are comparable, which would allow the reviewer to interpret a finding of similar (numerical) effects across designs as an indication of similar true treatment effects.

Proposed Policy: In some syntheses the differences in available indices and requisite assumptions will be large enough that results from the various subsets of studies in the synthesis will obviously not be comparable. In such cases we recommend that reviewers do not attempt to summarize across the designs, but rather include both sets of results in the review and report their results separately. In other cases, where comparability of the effects from different designs is conceivable (perhaps with certain assumptions), the reviewer may wish to summarize across designs. However, the assumptions underlying the comparison should be made explicit, and the reviewer should critically examine the data for the possibility of design-related differences in effects. Further, when such comparisons are made, the type of design should be tested as a moderator variable and separate results should be reported.

Furthermore, while the primary focus of C2 is on matters directly related to research

cumulation, activities that lead to improvements in the design of future research studies can be seen as auxiliary to that focus and certainly not contrary to it. C2 therefore takes the position that the study and careful application of methods of cross-design synthesis is consistent with the goals of C2.

6. What should be the role of C2's Social, Psychological, Educational and Criminological Trials Register (SPECTR) in supporting or informing the statistical research that might be done in the Campbell context?

The Campbell Collaboration SPECTR Database consists of entries for over 10,000 possibly randomized studies in social, psychological, educational and criminological research. While it is not a true representative (probability) sample of any domain of research, the SPECTR database can nevertheless be a rich source of information for methodological studies. It could be used as a sampling frame for studies that examine various questions about methodology for systematic reviews, much as the Cochrane Library has been used. However an important caveat is that SPECTR, a database of randomized trials, cannot provide information about nonrandomized studies, but a corresponding database of nonrandomized trials could obviously do so.

SPECTR can be used to *estimate normative methodological practice* in relevant domains of research that use, in the main, randomized experiments. Specifically, it could be used to study what kinds of designs and analyses have been most frequently used in studies that are likely candidates for a C2 synthesis. For example it could answer questions such as "How likely are ANCOVA designs?" or "How often will we need to deal with results reported as odds ratios?" or "How likely are we to encounter designs from which commensurable effect sizes cannot be computed for most studies?" Such information is relevant for developing analysis plans and coding manuals for syntheses that include information about handling the kinds of designs and reporting strategies that are likely to be encountered. It is also relevant for determining the features that are essential in software used for C2 syntheses.

Three aspects of normative methodological practice that are especially important for planning syntheses are the typical within-study sample sizes, the numbers of studies in an area that might be the basis for a systematic review, and the degree of between-study heterogeneity in effects. Information on these aspects is necessary for statistical power analyses in meta-analysis (Hedges & Pigott, 2001). Similarly, sample size information about subgroups is essential to determine feasibility of subgroup analyses and estimate their statistical power.

Another aspect of normative practice is the kinds of subject populations, specific outcome measures, and treatment contexts that have been studied. These considerations are relevant to specifying the probable generalizability of syntheses. Here again SPECTR can be used to obtain insight about the range of situations faced in C2 syntheses.

SPECTR can also be used to *study reporting practices* that are relevant to statistical practice. For example, when a study reports statistics on differences in gains with no raw posttest information, but the desired effect-size metric is the mean difference standardized by raw posttest scores, the pretest-posttest correlation is needed to compute the effect-size estimate (McGaw & Glass, 1980). If no such correlation is reported (and cannot be deduced from what is reported) the pretest-posttest correlation has to be imputed. Studies using SPECTR could suggest whether such situations (and those requiring other imputations) are likely to occur in C2 syntheses. This information is important in making choices among effect-size metrics to be used, since some metrics may be more readily calculated from typically available data than others.

SPECTR can be used to collect information to help *create a better framework for imputing information* needed in some analyses but not reported in individual studies. For example, the pretest-posttest correlation needed to estimate change-score standard deviations in some designs is often not reported. However some studies do report these correlations. Studies of the distribution of values observed for various types of studies (e.g., types of pretests and posttests and intervals between them, for various constructs) could provide a better empirical basis for imputation and sensitivity analyses.

Similarly, when samples are clustered, for example in studies that sample intact clusters such as classes or schools, the ordinary standard deviation underestimates the standard deviation of an entire population. Taking this into account when computing the standard errors of the effect-size estimates (much like computation of design effects in sample surveys) requires the intraclass correlation (Rooney & Murray, 1996). The intraclass correlation is rarely reported, but evidence for imputation could be improved by collection of evidence from studies that do report this statistic.

SPECTR could also be used *to examine the association between the results obtained in syntheses and the methods used*. While important insights can be obtained from such work, it is critical to realize that these examinations are observational studies and therefore subject to many possible confoundings. In particular, observational studies of statistical methods are not a substitute for deductive theory. For example, comparing the results of fixed- and random-effects analyses in studies that report both might show that results almost always differ substantially, but this could be a consequence of the fact that authors report both analyses precisely when they differ. Deductive analysis shows that they need not differ substantially and shows precisely when they do or do not lead to different results

(Hedges & Vevea, 1998).

Proposed Policy: The Steering Committee should endorse the use of SPECTR for research on normative methodological and reporting practice in relevant research domains, improving information for imputation in effect-size computation, and studying associations between synthesis methods and results.

References

- Becker, B.J. (1987). Applying tests of combined significance in meta-analysis. *Psychological Bulletin*, *102*, 164-171.
- Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, *60*, 373-417.
- Becker, B.J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*, *17*, 341-362.
- Becker, B.J. (1995). Corrections to "Using results from replicated studies to estimate linear models". *Journal of Educational Statistics*, *20*, 100-102.
- Becker, B. J. (2000). Multivariate meta-analysis. In H.E.A. Tinsley & S. Brown (Eds.), *Handbook of applied and multivariate statistics and mathematical modeling*. San Diego: Academic Press.
- Boruch, R.F., & Terhanian, G. (1998). Cross-design synthesis. In A. J. Reynolds & H. J. Walberg (Eds.), *Advances in educational productivity* (Vol. 7, pp. 59-85). Greenwich, CT: JAI Press, Inc.
- Cook, D.J., Sackett, D.L., & Spitzer, W.O. (1995). Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on meta-analysis. *Journal of Clinical Epidemiology*, *48*, (1), 167-171.

- Cooper, H. (1988). The structure of knowledge synthesis: A taxonomy of literature reviews. *Knowledge in Society*, 1, 104-126.
- Cooper, H. M. (1989). *Integrating research*. Newbury Park, CA: Sage Publications.
- DerSimonian, R., & Laird, N. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review*, 53, 1-15.
- Droitcour, J.A., Silberman, G., & Chelimsky, E. (1993). Design synthesis. *International Journal of Technology Assessment in Health Care*, 9(3), 440-449.
- Dunlap, W.P., Cortina, J.M., Vaslow, J.B., & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, 1, 170-177.
- Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. Pages 339-356 in H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage.
- Greenhouse, J.B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. Pages 383-398 in H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis*, Russell Sage Foundation, New York.
- Greenwald, R., Hedges, L.V., & Laine, R.D. (1996). The effect of school resources on student-achievement. *Review of Educational Research*, 66(3), 361-396.
- Haddock, C. K., Rindskopf, D. & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*, 3, 339-353.
- Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests.

Psychological Bulletin, 117, 167-178.

Hedges, L.V. (1994). Fixed effects models. Pages 285-300 in H. Cooper and L.V. Hedges (Eds.) *The handbook of research synthesis*. New York: Russell Sage Foundation.

Hedges, L. V. & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin*, 88, 359-369.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6, 203-217.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods*, 6, 486-504.

Kalaian, H., & Raudenbush, S.W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, 1, 227-235.

Light, R. J., & Pillemer, D. (1984). *Summing up*. Cambridge: Harvard University Press.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. Pages 439-453 in H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis*, Russell Sage Foundation, New York.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

McGaw, B., & Glass, G. V (1980). Choice of metric for effect size in meta-analysis. *American*

Educational Research Journal, 17, 325-337.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., et al; (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *The Lancet*, 354 (9193), 1896-1900.

Morris, S.B., & DeShon, R.P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7, 105-125.

Pigott, T.D. (1994). Methods for handling missing data in research synthesis. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage.

Raudenbush, S. W., Becker, B. J., & Kalaian, S. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, 102, 111-120.

Rooney, B. L., & Murray, D. M. (1996). A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis. *Health Education Quarterly*, 3,48-64.

Rosenthal, R. (1994). Parametric measures of effect size. Pages 231-244 in H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis*, Russell Sage Foundation, New York.

Stroup, D.F., Berlin, J.A., Morton, S., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Journal of the American Medical Association*, 283 (15), 2008-2012.

Timm, N.H. (1999). Testing multivariate effect sizes in multiple-endpoint studies. *Multivariate Behavioral Research*, 34 (4), 457-465.

Appendix A: Books on Meta-analysis and Meta-Analysis Examples

Books

Cook, T.D., Cooper, H.M., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A., & Mosteller, F. (1992). *Meta-analysis for explanation: A casebook*. New York: Russell Sage Foundation.

Cooper, H.M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Beverly Hills, CA.: Sage Publications.

Cooper, H.M., & Hedges, L.V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Eddy, D.M., Hasselblad, V., & Schachter, R. (1992). *Meta-analysis by the confidence profile method: The statistical synthesis of evidence*. San Diego: Academic Press.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA.: Sage Publications.

Hedges, L.V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias research findings*. Beverly Hills, CA.: Sage Publications.

Hyde, J.S., & Linn, M.C. (1986). *The psychology of gender: Advances through meta-analysis*. Baltimore: Johns Hopkins Press.

Light, R.J., & Pillemer, D.B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Mullen, B. (1989). *Advanced BASIC meta-analysis*. Hillsdale, NJ: Erlbaum.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications.

Wachter, K.W., & Straf, M.L. (Eds.). (1990). *The future of meta-analysis*. New York: Russell Sage Foundation.

Wolf, F.M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA.: Sage Publications.

Examples of Meta-Analysis

Examples of syntheses designed to estimate the average effect of a treatment

Beck, R., & Fernandez, E. (1998). Cognitive-behavioral therapy in the treatment of anger: A meta-analysis. *Cognitive Therapy and Research*, 22(1), 63-74.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using o^2 and

d. *American Psychologist*, 26, 892-901.

Examples of syntheses designed to determine if any study showed an effect

Astin, J.A., Harkness, E., & Ernst, E. (2000). The efficacy of "distant healing": A systematic review of randomized trials. *Annals of Internal Medicine*, 132 (11), 903-910.

Rosenthal, R. & Rubin, D. B. (1978b). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377-386.

Example of synthesis designed to study variation

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529-540.

Examples of syntheses designed to determine which variation of a treatment works best

Smith, M.L., & Glass, G. V (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-760.

Swanson, H.L. & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research*, 68 (3), 277-321.

Appendix B: Statistics that should be Reported in Research Syntheses

Conclusions about the average effect size

1. Fixed-effects models

- a. The value of each study's fixed effect-size estimate and standard deviation.
- b. The fixed-effects weighted mean effect size across studies and its 95% confidence interval.
- c. Recommended: Influence statistics - the weight for each study's effect size as a proportion of the total weight of the study.

2. Random-effects models

Same as above using the random-effects model

Conclusions about the variability or consistency of an effect

1. Fixed-effects models

- a. Confidence interval plots
- b. Homogeneity statistic, its associated degrees of freedom, and significance test

2. Random-effects models

- a. Confidence interval plots
- b. Variance-component estimate
- c. Test of variance component, its associated degrees of freedom and significance level.

Conclusions about the relationship between effect sizes and study characteristics

A. Analysis of variance models

1. Fixed-effects models

- a. Weighted mean effect size for each group, and its 95% confidence interval
- b. Homogeneity statistic within each group, its associated degrees of freedom, and significance test
- c. Between-groups homogeneity test, its associated degrees of freedom, and significance test

2. Random-effects models

- a. Weighted mean effect size for random-effects model for each group and its 95% confidence interval
- b. Variance-component estimate for each group
- c. Test of variance component, its associated degrees of freedom and significance level within each group.
- d. Between-groups homogeneity test, its associated degrees of freedom, and significance test

B. Regression models of effect size

1. Weighted least squares analysis using appropriate weights depending on fixed- or random-effects assumption
2. Adjusted standard errors for the regression coefficients and the 95% confidence interval for the regression coefficients (Hedges & Olkin, 1985)
3. Weighted residual sum of squares, degrees of freedom and significance level, for both fixed- and random-effects models.
4. Examination of standardized residuals for both fixed- and random-effects models.

Appendix C: Statistical Information that should be Coded

Statistical analysis of study

1. Effect-size estimates and their standard errors (or variances)
2. Type of statistical analysis
 - a. Significance level of statistical test (p value or p-value range)
 - b. Direction of finding in statistical test
 - c. Is test directly measuring desired effect?
3. Amount of missing data - some measure of the number of cases included in the statistical test versus the number in the original sample
4. Point in time when the measure was administered (in relation to treatment interval)
5. If the outcome was measured at other times (e.g., multiple follow-ups), what were they?
6. Reliability of the measure
7. Range restriction of the measure
8. In pretest-posttest designs, the correlation between the pretest and the posttest. If there is a correction for design effects of clustering, what design effect or intraclass correlation was used?

Characteristics of the effect size

1. Type of effect size - posttest measures only, ANCOVA-adjusted means, differences in pretest to posttest gains, correlations, odds ratios, etc.
2. Type of adjustment used for pretest nonequivalence, if relevant
3. Computation procedure used - effect size computed from means and standard deviations, from t-test value, from sums of squares in multi-factor ANOVA, from significance levels, etc.
4. Was any imputation necessary to compute this effect size (e.g., imputed pretest-posttest correlation or intraclass correlation)?

(1)The authors are members of the statistics group, a registered C2 entity.

(2)The order of authorship was determined alphabetically.

(3)Clearly a database of comparative trials cannot be used to make inferences about the nature of related studies using other designs.

[1] The authors are members of the statistics group, a registered C2 entity.

[2] The order of authorship was determined alphabetically.

[3] Clearly a database of comparative trials cannot be used to make inferences about the nature of related studies using other designs.