

Campbell Collaboration  
Research Design Policy Brief<sup>[1]</sup>

Prepared for the Campbell Collaboration Steering Committee by:

William Shadish

University of Memphis

and

David Myers

Mathematica Policy Research

Campbell Collaboration  
Research Design Policy Brief

Executive Summary

This Brief addresses the following key question for Campbell Collaboration (C2) reviews: *What should be C2 policy concerning acceptable methodologies used in primary studies when a systematic review concerns the effectiveness of an intervention?* The Brief:

1. identifies the key issues that are confronted by C2 systematic reviewers who find a variety of study designs in their literature;
2. outlines possible ways to represent this diversity in their work;
3. proposes agreed-upon guidelines that C2 may wish to promulgate; and
4. provides exemplars that demonstrate how these guidelines might be implemented in practical ways.

The heart of the Brief is a set of eight key issues, along with proposals for C2 policies for each. A summary of those issues and a summary of the proposals is:

1. Should the C2 reference database be limited to randomized experiments?  
Proposal: The C2 database(s) should not be limited to randomized experiments.
2. If nonrandomized studies are included, should one C2 reference database include  
both randomized and nonrandomized studies, or should separate databases be constructed for the two kinds of studies?  
Proposal: C2 should maintain two databases, one for randomized experiments and one for nonrandomized studies.  
Proposal: C2 should address how to handle the borderline case of haphazard assignment, that is, assignment methods that appear to be functionally random in how they distribute bias over conditions.  
Proposal: Group or cluster randomized trials with discrepancies between unit of assignment and unit of analysis, which lead to incorrect calculation of standard errors, or with small samples of aggregate units should not be combined with other randomized experiments.
3. What searchable fields should be available for users to identify research designs in

the database(s)?

Proposal: The searchable fields initially included to identify research designs in the databases should be kept to the minimum that potential reviewers might need to select studies for their review. The list should include:

1. Randomized Design (include group randomized designs here unless they meet criteria for the following code)
  2. Group Randomized Design with Discrepant Units of Analysis or with Inadequate Number of Aggregate Units Assigned to Conditions
  3. Quasi-Experiment: Interrupted Time Series Design
  4. Quasi-Experiment: Regression Discontinuity Design
  5. Quasi-Experiment: Nonequivalent Comparison Group Design
  6. Case Control Design
  7. Other Designs
4. Should C2 systematic reviews or quantitative syntheses be limited to randomized experiments?

Proposal: C2 systematic reviews may include either randomized and nonrandomized experiments, or both, as part of the process of judging the nature of the available evidence on a question. However, with some exceptions described in more detail in the Brief, C2 reviews should not include a quantitative synthesis of these studies unless randomized experiments are available within the body of evidence to be synthesized. Such quantitative syntheses need not be limited to randomized experiments.

5. If nonrandomized studies are included in C2 quantitative syntheses, what procedures should be required or recommended for determining what nonrandomized designs are legitimate for inclusion?

Proposal: Where both randomized and nonrandomized experiments are included, C2 quantitative syntheses *must* separate estimates of intervention effects for randomized versus nonrandomized studies in important analyses. The Brief contains more specific recommendations about how to treat different kinds of designs in many other respects, including differentiating among better and worse randomized experiments, or better and worse quasi-experiments.

6. Should a standard set of design coding categories be used in C2 systematic reviews or quantitative syntheses?

Proposal: C2 should develop a standard set of codes to be used in reviews and syntheses for purposes of coding study design and related features. Some such codes are presented and discussed in the Appendix to this Brief.

7. Of these design codes, which should C2 reviews and syntheses be required to code

and which should be recommended (optional)?

Proposal. Specific recommendations are made for which codes are required and which are optional in the text of the Brief.

8. Should C2 consider grading the quality of either randomized or nonrandomized experiments, or both, as to level of defensibility of results?

Proposal. Aspects of methodology that are related to the validity of a study's conclusions should be assessed individually rather than being summed into a total quality score.

More details on all these matters, including more detailed proposals, are described in the text of this Brief.

## Campbell Collaboration

### Research Design Policy Brief

#### *Preamble*

The methods we use in science should follow from the questions we ask. Because Campbell Collaboration (C2) systematic reviews examine a variety of different questions, C2 reviews will include studies that use different kinds of methodologies.

The purpose of this Policy Brief is to address one important issue: *What should be C2 policy concerning acceptable methodologies used in primary studies when a systematic review concerns the effectiveness of an intervention?* The material that follows is intended to apply only to C2 reviews on intervention effectiveness, and not to C2 reviews that might address other interesting questions such as treatment implementation, needs assessment, or rich description of participant experiences. Moreover, the policies described in this brief are intended to apply only to that part of a C2 systematic review that summarizes the effects of an intervention. For example, nothing in the material that follows is intended to preclude the use of any scientific work to inform C2 systematic reviews about how well an intervention was implemented. Indeed, a systematic review may be better for including the latter kind of information. Finally, C2 intends this to be an evolving document, with periodic changes anticipated as experience with C2 systematic reviews accumulates.

#### *Procedure*

The Research Design Policy Brief was initiated by the Campbell Collaboration Methods Group, which assigned the task of developing the Brief to the co-conveners of the Quasi-Experimental Design Group (William Shadish, David Myers). Shadish and Myers developed an initial draft in December 2001 based on a review of the literature, and on consultations with experienced meta-analysts and methodologists. That draft was then circulated to a panel of five outside experts (Tom Cook, Harris Cooper, Diana Elbourne, Mark Lipsey, Paul Rosenbaum, Donald Rubin). Some of their comments were incorporated directly into the draft, or reflected endorsements of the draft. Other comments reflected substantial disagreements with aspects of the draft that could not be incorporated completely into the current draft; these are quoted (anonymously) in a footnote to the relevant text. Presumably these footnotes will eventually be deleted once C2 decides the pertinent issues. The authors also sought input on the eight key issues in this Brief at the second annual C2 colloquium (February 21-22, 2002), and that input was also incorporated into the Brief. The result of this process is the present document, which is presented for further criticism and comment by a wider audience.

#### *Introduction*

Systematic reviews of the effects of interventions can rely on primary studies that use a wide variety of designs. These designs may engender different patterns of threats to internal validity and thereby permit causal inferences with different levels of certainty. Given that Campbell Collaboration (C2) systematic reviewers are likely to encounter a

variety of designs, in this policy brief we attempt to:

1. identify the key issues that are confronted by C2 systematic reviewers who find a variety of study designs in their literature;
2. outline possible ways to represent this diversity in their work;
3. identify agreed-upon guidelines that C2 wishes to promulgate; and
4. provide exemplars that demonstrate how these guidelines might be implemented in practical ways.

### *Systematic Reviews and Quantitative Syntheses*

In the material that follows, the distinction between systematic reviews and quantitative syntheses is crucial to the recommendations. Not all systematic reviews will proceed to do a quantitative synthesis. For example, a reviewer may find no studies on a question of interest, or may find no studies of sufficient quality to synthesize quantitatively. Therefore, criteria for what might be included in a systematic review are necessarily looser than criteria for what might be included in a quantitative synthesis.

### **Key Issues**

The panel, in collaboration with the C2 Steering Committee, identified the following eight key issues for consideration in a C2 Policy Brief on Research Design:

1. Should the C2 reference database be limited to randomized experiments?
2. If nonrandomized studies are included, should one C2 reference database include both randomized and nonrandomized studies, or should separate databases be constructed for the two kinds of studies?
3. What searchable fields should be available for users to identify research designs in the database(s)?
4. Should C2 systematic reviews or quantitative syntheses be limited to randomized experiments?
5. If nonrandomized studies are included in C2 quantitative syntheses, what procedures should be required or recommended for determining what nonrandomized designs are legitimate for inclusion?
6. Should a standard set of design coding categories be used in C2 systematic reviews and quantitative syntheses?

7. Of these design codes, which should systematic reviewers be required to code and which should be recommended (optional)?
8. Should C2 consider grading the quality of either randomized or nonrandomized experiments, or both, as to level of defensibility of results?

In the sections that follow we (a) describe these issues in more detail, (b) identify possible solutions, and (c) suggest guidelines for how the issues might be handled in C2 reviews.

#### 1. Should the C2 reference database be limited to randomized experiments?

Properly implemented randomized experiments address the problem of selection biases by using random assignment to ensure that groups of participants (or other units) are probabilistically equivalent prior to the administration of a treatment. In research designs that do not use random assignment, the units who receive different interventions may differ from each other systematically (i.e. other than by chance), and hence the effects of the intervention may be difficult to disentangle from these existing (and often unknown) differences. In addition, properly implemented randomized experiments can yield estimates of the effects of interventions that have very desirable statistical and inferential properties. Specifically, effect estimates from randomized experiments are consistent, that is, they converge on population parameters as sample size increases. Also, the estimates are accompanied by confidence intervals that have known probabilities of containing the population parameter, and that are often narrower than intervals from some other designs. No other design is known to have all these properties. Furthermore, given the general high regard with which randomized experiments are held among methodologists, policymakers and other users often give considerable credibility to results from randomized experiment. For such reasons, the Campbell Collaboration's sister organization, the Cochrane Collaboration, has tended to limit its database to randomized experiments. This decision has received widespread support in the science and user communities. All these reasons support limiting the C2 reference database to randomized experiments.

However, other reasons suggest that C2 give nonrandomized studies some place in its database(s). First, even the Cochrane Collaboration has not adhered to this limitation strictly. For example, the Cochrane Library has registered reviews that consist of case control designs (Thompson & Rivara, 2001; Thompson, Rivara, & Thompson, 2000; Stroup et al., 2000) and interrupted time series quasi-experiments (Grilli, Freemantle, Minozzi, Domenighetti, & Finer, 2000).

Second, randomized experiments tend to be plentiful in some fields, especially the medical and public health fields on which the Cochrane Collaboration focuses, but are less plentiful in many of the social science fields that are the focus of C2. For some questions in C2 fields there may never be enough randomized studies to warrant a research synthesis. Thus, limiting the C2 database to randomized experiments only might significantly curtail the topics about which C2 can mount reviews.

Third, nonrandomized studies may offer some kinds of information less widely available in randomized experiments. For example, nonrandomized designs may examine intervention features, settings, or kinds of service recipients that are less amenable to random

assignment, and that therefore would be under-represented in a database limited only to randomized experiments.

Fourth, nonrandomized experiments themselves vary enormously in quality. For example time series and regression discontinuity designs tend to yield generally strong causal inferences while posttest-only studies with no control group rarely do. To exclude all nonrandomized designs discourages researchers from conducting high quality nonrandomized experiments.

Fifth, randomized experiments also vary in quality. Sometimes, they suffer from extremely high differential attrition rates, low sample sizes, and poor treatment implementation. These flaws can make their results far less helpful than better conducted randomized experiments or even than some high quality nonrandomized experiments.

Finally, while the synthesis of results across studies will not cancel out consistent and pervasive design flaws, to the extent that the strengths of some studies, randomized or not, compensate for the weaknesses of others (and vice versa), clear patterns that emerge across flawed bodies of evidence can be informative.

Proposal: The C2 database(s) should not be limited to randomized experiments. They should make some provision for possible inclusion of nonrandomized studies. This policy should be coupled with clear guidelines about how different designs should or should not be combined in C2 reviews. Useful database codes should be available to help reviewers clarify the warrant for causal inferences in nonrandomized studies, and exemplars should be provided for how these interpretations might be carried out. This strategy will allow C2 (a) to take maximum advantage of the full set of information available across a wide array of study designs, (b) encourage the conduct of high quality nonrandomized experiments, and (c) contribute to the development of empirical knowledge about the conditions under which nonrandomized studies might provide accurate estimates of the effects of interventions.

## **2. If nonrandomized studies are included, should a single C2 reference database include both randomized and nonrandomized studies, or should separate databases be constructed for the two kinds of studies?**

Assuming that any database would contain identifiers for the kind of design used by each study in the database, then this decision may have more rhetorical than logical import. That is, under either option, the database would still have a record of the kind of study design used, and those retrieving studies from the database would have access to the same design information in either case. However, maintaining two databases makes a clear statement to reviewers and to other C2 stakeholders about the special significance that C2 attaches to random assignment as a design feature.

Proposal: C2 should maintain two databases, one for randomized experiments and one for nonrandomized studies<sup>[2], [3]</sup>.

Proposal: C2 should address how to handle the borderline case of haphazard assignment, that is, assignment methods that appear to be functionally random in how they distribute bias over conditions. Examples include (a) clients being assigned to treatment or control in

alternating weeks (Coche & Flick, 1975), and (b) alternating assignment of clients to treatment or control until treatment slots were filled with all subsequent applicants placed in control (Endicott & Endicott, 1964).

In general, C2 systematic reviewers should not code such studies as using random assignment unless the case is clear that the haphazard mechanism was random in practice (e.g., Coche & Flicke, 1975). When doubt exists (e.g., Endicott & Endicott, 1964), these studies should be coded as nonrandom assignment, or they should be placed in a separate category called "haphazard assignment" that can be compared to other assignment mechanisms during subsequent analyses.

**Proposal:** Group or cluster randomized experiments<sup>[4]</sup> refer to studies in which aggregate units such as families, worksites, or schools are randomly assigned to conditions (Feng, Diehr, Peterson & McLerran, 2001). When a reasonably large number of aggregate units is randomly assigned to conditions, and when the data are properly analyzed, results from such studies should be treated as randomized experiments (though there can be some question about exactly how to calculate effect size estimates from such studies, e.g. Murray et al., 1994). However, sometimes random assignment takes place at a level different from the level employed in the data analysis. In other cases very few or only one aggregate unit is included in each condition (Varnell, Murray & Baker, 2001). For example, classes or schools may be randomly assigned to conditions but students are the unit of analysis, or only one school per condition may be randomized. While such studies are formally randomized experiments, the inappropriate specification of the unit for statistical analysis can create situations where incorrect standard errors for impact estimates are computed, which may lead to incorrect conclusions about the effectiveness of treatments. Furthermore, small number of units can thwart the goals of randomization in ways that rarely occur when individuals are randomized to conditions. Consequently, group randomized trials with such misanalyses or with small samples should not be combined with other randomized experiments.

### **3. What searchable fields should be available for users to identify research designs in the database(s)?**

Having a comprehensive set of fields in the C2 database could potentially save time and effort for C2 reviewers and could also standardize the coding of research design. However, comprehensive fields could also create a complex logistical process for people entering studies into the C2 database, and might waste resources if some of the codes are not used by subsequent C2 reviewers.

**Proposal:** The searchable fields initially included to identify research designs in the databases should be kept to the minimum that potential reviewers might need to select studies for their review. The list should include:

1. Randomized Design (include group randomized designs here unless they meet criteria for the following code)
2. Group Randomized Design with Discrepant Units of Analysis or with Inadequate

## Number of Aggregate Units Assigned to Conditions<sup>[5]</sup>

3. Quasi-Experiment: Interrupted Time Series Design<sup>[6]</sup>
4. Quasi-Experiment: Regression Discontinuity Design
5. Quasi-Experiment: Nonequivalent Comparison Group Design
6. Case Control Design
7. Other Designs

### 4. Should C2 systematic reviews or quantitative syntheses be limited to randomized experiments?

The issues here are largely similar to those covered earlier concerning whether the C2 database should be limited to randomized experiments. Randomized experiments often have greater internal validity and credibility associated with effect estimates, while nonrandomized studies can provide important information, especially for review topics for which it might prove difficult to locate large numbers of randomized experiments. In addition, recall that even the Cochrane Collaboration has not strictly limited either its reviews or its quantitative syntheses to randomized experiments, suggesting that some flexibility on this issue may be appropriate, at least occasionally. Finally, the inclusion of both randomized and nonrandomized studies in C2 reviews and syntheses may serve (a) an educative function by informing users and the general public about the importance of random assignment when randomized and nonrandomized studies applied to the same treatment yield different results, and (b) a scientific function by clarifying conditions under which nonrandomized studies might yield accurate results.

Proposal: C2 systematic reviews may include either randomized and nonrandomized experiments, or both, as part of the process of judging the nature of the available evidence on a question. However, with some exceptions described shortly, C2 reviews should not include a quantitative synthesis of these studies unless randomized experiments are available within the body of evidence to be reviewed. C2 quantitative syntheses need not be limited to such experiments<sup>[7], [8], [9]</sup> but exceptions to this rule must be well-justified<sup>[10]</sup>. Examples of exceptions might include the following: (a) quantitative syntheses of studies that use either interrupted time series designs or regression discontinuity designs, given that these designs generally yield high quality causal inferences; (b) quantitative syntheses whose aims do not include producing strong causal inferences, such as syntheses of case control studies that aim to generate causal hypotheses or to describe the implementation of a treatment, and (c) quantitative syntheses on topics for which it is generally understood that randomized experiments are rarely or never feasible, for example when there are ethical or practical reasons why randomization cannot occur, such as certain questions about effective neonatal care practices (e.g., Ozminkowski, Wortman & Roloff, 1989).

5. If nonrandomized studies are included in C2 quantitative syntheses, what procedures should be required or recommended for determining what nonrandomized designs are legitimate for inclusion?

Researchers who do quantitative syntheses have used very different approaches to how they treat randomized and nonrandomized studies. For example, some researchers combine both kinds of studies in all analyses; others test for differences in results associated with designs but combine them no matter what the results of the test were; and others test for design differences then use regression analyses to control for or take into account design and other methodological features when testing substantive moderators of effect. Given that some of these procedures are probably poor practice, there is benefit for C2 to have at least some recommended or default procedures for researchers. Though it is probably impossible to write simple and highly specific guidelines that cover all contingencies, the following proposals would provide C2 with some general guidelines that can be used to start discussion and generate experience.

Proposal: Where both randomized and nonrandomized experiments are included, C2 quantitative syntheses *must* report separate estimates of intervention effects for randomized versus nonrandomized studies in important analyses. Further, separate effect sizes should be provided when multiple distinguishable classes exist of randomized designs (e.g., with and without high differential attrition) and of nonrandomized designs (e.g., nonequivalent comparison group designs versus case control designs). These results should be reported even if estimates based on different designs are judged to be similar or are not statistically significantly different from each other. Further, subsequent analyses pertaining to moderators of the effects of the intervention should not combine results when initial tests indicate that results from the two kinds of designs differ; indeed, they should be cautious about combining them even if they do not differ, for further differences may appear during subsequent moderator analysis. These moderators should include both methodological and substantive variables, such as different kinds of treatments or settings. Given that the use of randomization may be correlated with such substantive variables, the goal of all these breakdowns is to understand both the methodological and substantive factors that might contribute to study results in both randomized and nonrandomized experiments, and whether they act similarly or differently across designs.

Even when initial results from the two kinds of experiments are similar, C2 reviewers should be cautious about merging them in subsequent analyses. A lack of overall difference does not guarantee a lack of differences when studies are broken down into subsets. However, results from randomized and nonrandomized studies may be combined in some analyses. For example, combining might be appropriate when it is clear from multiple analyses that results are the same no matter which kinds of experiments are included. Or, sufficiently large numbers of studies may exist to allow univariate and multivariate exploration of why differences in results might emerge from the two kinds of experiments. The latter explorations are strongly encouraged in order to build theory about the conditions under which nonrandomized studies can yield useful and accurate estimates of treatment effects.

6. Should a standard set of design coding categories be used in C2 systematic reviews or quantitative syntheses?

C2 should guide reviewers about how to code pertinent aspects of research design. This will assist less experienced reviewers and will help subsequent users who wish to use reviews in studies of methodology. This guidance should include at least some design codes that are used in all relevant C2 reviews.

**Proposal:** C2 should develop a standard set of codes to be used in reviews for purposes of coding study design and related features. Rationales should be provided for (a) why the code is suggested, (b) how it should be interpreted, (c) whether the code is intended to be used for all kinds of designs or just for some of them, and (d) how it can be used in analysis. References suggesting that the code may be related to effect size should also be provided. Although this task is started in the present paper (see Appendix), much more work remains. Moreover, C2 encourages reviewers to explore a wide variety of other possible codes that may aid understanding of how research methodology may influence the results of systematic reviews and quantitative syntheses. 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 methodological 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. In this way, C2 coding procedures can evolve and improve over time.

### **7. Of these design codes, which should C2 reviews and quantitative syntheses be required to code and which should be recommended (optional)?**

If the set of proposed codes is large, C2 reviewers may balk at coding them all. Hence, the codes may benefit from being categorized into smaller subsets, perhaps hierarchically ordered by importance in C2 systematic reviews.

**Proposal.** All C2 systematic reviews and quantitative syntheses must use codes 1, 6, and 7 in the Appendix in order to provide a general category for the design used, and to allow computation of weights for effect sizes. Systematic reviews and quantitative syntheses that mix both randomized and nonrandomized designs must use code 2. Systematic reviews and quantitative syntheses that mix both randomized and nonrandomized designs may optionally use the remaining codes in analyses that either (a) explore reasons that might explain discrepancies in effect sizes among different kinds of designs, or (b) use design features as covariates in regression analyses to attempt to control statistically for them.

### **8. Should C2 consider grading of either randomized or nonrandomized experiments, or both, as to level of defensibility of results?**

Many researchers who have reviewed outcome research either narratively or using meta-analytic techniques have created scales for grading the "quality" of studies. Their motivation is to take into account the widely-shared experience of researchers that studies can differ greatly in the validity of their inferences. Certainly, C2 reviewers should weight conclusions based on the strength of inferences permitted by the characteristics of the studies forming the evidential base.

For this reason, C2 encourages the exploration and development of mechanisms for understanding and improving the quality of research designs. However, a variety of concerns lead to significant reservations about whether C2 should endorse the routine use of any scale that results in a single number to represent study quality.

Scales that result in a single number representing design quality tend to combine multiple items into a total score that subsumes a very wide array of methodological variables related to quality. These can include design, sample size, measurement reliability, and representativeness of participants, to name just a few research characteristics. Clearly, these diverse items do not assess the same kind of "quality." Rather, some assess internal validity, some statistical conclusion validity, some construct validity, and some external validity. Thus, global quality assessments result in studies with decidedly different validity characteristics emerging with identical total scores. Lumping such diversity together may yield a total score of questionable utility or validity. Further, such scales are rarely well-developed psychometrically, with no evidence of factorial structure to justify scoring, or even evidence of simple internal consistency reliability. The conclusion reached about study quality, and the conclusion reached by the review itself, can differ considerably depending on which quality scale is used (Jüni, Witschi, Bloch & Egger, 1999). Finally, global decisions about what makes a study good or bad are fraught with difficulty. There is ample evidence that even the most sophisticated researchers can disagree about the dimensions that define quality and how these dimensions apply to particular studies. The effect of research design features on study outcomes is an empirical question. Low-inference, operational details of studies can be examined empirically for their relation to outcomes. Then, if studies with more desirable features produce results different from other studies, inferences about the literature can be adjusted accordingly.

Proposal. Aspects of methodology that are related to the validity of a study's conclusions should be assessed individually rather than being summed into a total quality score. C2 encourages further development of multi-dimensional quality scales as an important part of reviews. C2 reviewers should pay particular attention to (a) clarifying the kind of quality they believe the scale assesses, (b) providing psychometric data about the reliability, factors structure, and validity of the scales where multiple items scales are used, and (c) exploring how conclusions about the relationship between quality and outcome would vary depending on whether the total score or the individual items were used.

*An Annotated Bibliography of Exemplars of How to Include Both Randomized and Nonrandomized Studies in Systematic Reviews and quantitative syntheses:*

A number of previous meta-analyses have explored differences between randomized and nonrandomized experiments. They do so in different ways and illustrate some productive ways this can be done. These exemplars are meant to be illustrative rather than exhaustive, and we encourage C2 reviewers to explore this issue in other ways, as well.

Heinsman, D.T., & Shadish, W.R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods*, 1, 154-169. [see also Shadish, W.R., & Heinsman, D.T. (1997). Experiments versus quasi-experiments: Do you get the same answer? In W.J. Bukoski (Ed.), *Meta-Analysis of Drug Abuse Prevention Programs* (NIDA Research Monograph,

DHHS Publication No. (ADM) 97-170) (pp. 147-164). Washington DC: Superintendent of Documents.]

Shadish, W.R., & Ragsdale, K. (1996). Random versus nonrandom assignment in psychotherapy experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64*, 1290-1305.

These two meta-analyses were both methodological rather than substantive. They were aimed solely at understanding differences between randomized and nonrandomized experiments, rather than answering a particular question about treatment effectiveness. They used nearly the same coding schemes applied to different sets of randomized and nonrandomized experiments. They also used both categorical and regression analyses that explore why differences between these designs may have emerged.

Shadish, W.R., Matt, G.E., Navarro, A.M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis, *Psychological Bulletin, 126*, 512-529.

This was a substantive meta-analysis that mixed randomized and nonrandomized experiments. All results were reported separately for randomized experiments only, and for all studies pooled together. Primary analyses were random effects regression equations that statistically controlled for multiple covariates simultaneously. A separate section of the results was devoted to exploring discrepancies between results from randomized and nonrandomized experiments. In that exploration, codes for pretest effect size and self- versus other-selection proved crucial, and were related in a plausible way to a substantive explanation for how self-selection might have caused a bias.

Ozminkowski, R.J., Wortman, P.M., & Roloff, D.W. (1989). Inborn/outborn status and neonatal survival: A meta-analysis of non-randomized studies. *Statistics in Medicine, 7*, 1207-1221.

This was a substantive meta-analysis of 19 nonrandomized experiments comparing survival rates of low birth weight infants cared for in neonatal intensive care units to those not cared for in such units. The authors identified two possible kinds of selection bias that might occur in these studies, and then geared a number of analyses and discussion to assessing the plausibility of these two biases. Given the ambiguities caused by these selection biases, the Discussion section of this article provides a useful model for how to appropriately qualify confidence in results from these studies.

Wortman, P.M., & Bryant, F.B. (1985). School desegregation and black achievement: An integrative review. *Sociological Methods & Research, 13*, 289-324.

This literature is dominated by nonrandomized experiments. The authors made a reasonable empirical case that a substantial portion of the selection bias in these quasi-experiments could be removed by subtracting the pretest effect size from the posttest effect size, and using the resulting difference in analyses.



## References

- Coche, E. & Flick, A. (1975). Problem-solving training groups for hospitalized patients. *Journal of Psychology*, 91, 19-29.
- Endicott, N.A., & Endicott, J. (1964). Prediction of improvement in treated and untreated patients using the Rorschach prognostic rating scale. *Journal of Consulting Psychology*, 28, 342-348.
- Feng, Z., Diehr, P., Peterson, A., & McLerran, D. (2001). Selected statistical issues in group randomized trials. *Annual Review of Public Health*, 22, 167-187.
- Grilli, R., Freemantle, N., Minozzi, S., Domenighetti, G., & Finer, D. (2000). Mass media interventions: Effects on health services utilization (Cochrane Review). *The Cochrane Library*, 3. Oxford: Update Software.
- Heinsman, D.T., & Shadish, W.R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods*, 1, 154-169.
- Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, 282, 1054-1060.
- Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181-1209.
- Murray, D.M., Rooney, B.L., Hannan, P.J., Peterson, A.V., Ary, D.V., Biglan, A., Botvin, G.J., Evans, R.I., Flay, B.R., Futterman, R., Getz, J.G., Marek, P.M., Orlandi, M., Pentz, M.A., Perry, C.L., & Schinke, S.P. (1994). Intraclass correlation among measures of adolescent smoking: Estimates, correlations, and applications in smoking prevention studies. *American Journal of Epidemiology*, 140, 1038-1050.
- Ozminkowski, R.J., Wortman, P.M., & Roloff, D.W. (1989). Inborn/outborn status and neonatal survival: A meta-analysis of non-randomized studies. *Statistics in Medicine*, 7, 1207-1221.
- Sacks, H.S., Chalmers, T.C., & Smith, H. (1982). Randomized versus historical controls for clinical trials. *The American Journal of Medicine*, 72, 233-240.
- Sacks, H.S., Chalmers, T.C., & Smith, H. (1983). Sensitivity and specificity of clinical trials: Randomized v historical controls. *Archives of Internal Medicine*, 143, 753-755.
- Shadish, W.R., & Heinsman, D.T. (1997). Experiments versus quasi-experiments: Do you get the same answer? In W.J. Bukoski (Ed.), *Meta-Analysis of Drug Abuse Prevention*

*Programs* (NIDA Research Monograph, DHHS Publication No. (ADM) 97-170) (pp. 147-164). Washington DC: Superintendent of Documents.

Shadish, W.R., Matt, G.E., Navarro, A.M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis, *Psychological Bulletin*, 126, 512-529.

Shadish, W.R., & Ragsdale, K. (1996). Random versus nonrandom assignment in psychotherapy experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology*, 64, 1290-1305.

Shadish, W.R., & Sweeney, R. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology*, 59, 883-893.

Spiegler, M. D., Cooley, E. J., Marshall, G. J., Prince II, H. T., Puckett, S. P., and Skenazy, J. A. (1976). A self-control versus a counterconditioning paradigm for systematic desensitization: An experimental comparison. *Journal of Counseling Psychology* 23, 83-86.

Swaen, G., Teggeler, O., & van Amelsvoort, L. (2001) False positive outcomes and design characteristics in occupational cancer epidemiology studies. *International Journal of Epidemiology*, 30, 948-954.

Thompson, D.C., Rivara, F.P., & Thompson, R. (2000). Helmets for preventing head and facial injuries in bicyclists (Cochrane Review). *The Cochrane Library*, 3. Oxford: Update Software.

Thompson D.C., & Rivara F.P. (2001). Pool fencing for preventing drowning in children (Cochrane Review). *The Cochrane Library*, 4. Oxford: Update Software.

Varnell, S., Murray, D.M., & Baker, W.L. (2001). An evaluation of analysis options for the one-group-per-condition design: Can any of the alternatives overcome the problems inherent in this design? *Evaluation Review*, 25, 440-453.

Weisburd, D., Lum, C.M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *The Annals*, 578, 50-70.

## Appendix 1:

## Possible Design Codes for Use in C2 Reviews

**1 Kind of Design**

(1) randomized experiment (include group randomized trials here if number of aggregates is adequate and if properly analyzed); (2) randomized experiment with units of analysis discrepancy or very small number of aggregate units (e.g., classrooms randomly assigned to conditions, but individuals treated as unit of analysis; or one school per condition); (3) quasi-experiment: interrupted time series; (4) quasi-experiment: regression discontinuity; (5) quasi-experiment: nonequivalent comparison group; (6) case control design; (7) Within-group comparison (i.e., pretest-posttest); (9) other (e.g., design that has both random assignment and self-selection into several groups).

Rationale (Required for all systematic reviews and quantitative syntheses): There is ample evidence in the meta-analytic literature that these designs can yield very different effect size estimates (e.g., Heinsman & Shadish, 1996; Lipsey & Wilson, 1993; Sacks, Chalmers & Smith, 1982, 1983; Shadish & Ragsdale, 1996; Shadish et al., 2000; Swaen, Teggeler & van Amelsvoort, 2001; Weisburd, Lum & Petrosino, 2001). This code allows quick identification of the general design used in the study. C2 should examine whether it would be desirable to make this code identical to that used in the C2 database(s).

**2 Randomization to the comparison made in the effect size**

(1) Participants were randomly assigned to between-participants treatment and comparison conditions for this effect size; (2) participants were haphazardly assigned (e.g., alternating order) to treatment and comparison conditions for this effect size; (3) participants were neither randomly nor haphazardly assigned to treatment and comparison conditions for this effect size; (4) this effect size is based on a within-participants comparison (e.g., a pretest-posttest design following the same participants over time); (9) Unknown.

Rationale (Required for systematic reviews and quantitative syntheses that include both randomized and nonrandomized designs): This code captures information that is not captured in the previous code for the following reasons. First, effect sizes are computed on a pair of means (a treatment "comparison"). However, it may be possible to compute effect

sizes on more than one comparison within a study (e.g., if a study has a treatment and two controls). Occasionally, one of those comparisons may be randomized and the other nonrandomized (Spiegler, Cooley, Marshall, Prince, Puckett, & Skenazy, 1976). Only a “comparison-specific” code can capture this possibility. Second, this code allows identification of studies using haphazard assignment.

### 3 Type of Comparison Condition

(1) Wait List Control Group; (2) No Treatment Control Group; (3) Placebo Control Group; (4) “Treatment as usual”; (5) An alternative treatment.

**Rationale** (Optional code that is recommended for use in exploration of design effects): This code has two justifications. First, the first three codes (and possible code (4) as well) identify those effect sizes pertaining to treatment-control comparisons, with code (5) being a treatment-treatment comparison. Meta-analysts often need to separate out these two kinds of comparisons, though there are some occasions in which they can be pooled. Second, within treatment-control comparisons, empirical findings suggest that effect sizes are smaller when treatments are compared to active control groups (placebo, treatment as usual) than when they are compared to passive control groups (no treatment, wait list) (Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996). Note, Code 10 below is an alternative coding of this same variable.

### 4 Initial Sample Size of Treatment Group

Number of units initially assigned to treatment group. For a randomized experiment, count all those units assigned to conditions even if they were later dropped, unless the dropped unit was later found to not meet inclusion criteria. For a nonrandomized experiment, count all those units assigned to start treatment. Be sure to count a unit the same way it is counted on the dependent variable being coded. For example, if families were the unit of analysis on the dependent variable, sample size should reflect the number of families, not the number of family members.

**Rationale** (Optional code that is recommended for use in exploration of design effects): This is necessary if the researcher wishes to compute attrition rates and analyze their impact on effect size<sup>[11]</sup>. That is, from this code and the next three, total attrition and differential attrition over conditions should be computed, and then examined for their relationship to effect size (highly recommended, including for meta-analyses of randomized experiments). Some previous meta-analytic evidence suggests attrition rates can be related to effect size (Shadish & Ragsdale, 1996).

## 5 Initial Sample Size of Comparison Group

Number of units initially assigned to comparison group. Use same rules as above.

Rationale (Optional code that is recommended for use in exploration of design effects): This is necessary if the researcher wishes to compute attrition rates and analyze their impact on effect size (highly recommended, including for meta-analyses of randomized experiments).

## 6 Sample Size of Treatment Group for This Effect Size

Number of units remaining in treatment group at the time the measure used in the effect size is taken.

Rationale (Required for all systematic reviews and quantitative syntheses): This is necessary if the researcher wishes to compute attrition rates and analyze their impact on effect size (highly recommended, including for meta-analyses of randomized experiments), and required so as to weight effect sizes by a function of sample size.

## 7 Sample Size of Treatment Comparison for This Effect Size

Number of units remaining in comparison group at the time the measure used in the effect size is taken.

Rationale (Required for all systematic reviews and quantitative syntheses): This is necessary if the researcher wishes to compute attrition rates and analyze their impact on effect size (highly recommended, including for meta-analyses of randomized experiments), and required so as to weight effect sizes by a function of sample size.

## 8 Matching, Blocking, Stratifying

Did the study use matching, blocking, or stratifying (before assigning participants to conditions for randomized experiments; or for quasi-experiments, in the process of creating the control group)? (1) matching/stratifying were used; (9) unknown.

**Rationale** (Optional code that is recommended for use in exploration of design effects): This should be coded separately from whether random assignment occurred or not because matching, blocking, or stratifying can be done either with random assignment or without it. Previous meta-analytic research indicates that this variable can predict effect size (Shadish & Sweeney, 1991), especially if it is used to help create more detailed design groupings (e.g., randomized experiments with low attrition, randomized experiments with high attrition, quasi-experiments with matching, etc.) (Shadish & Ragsdale, 1996). Though analysis of covariance is functionally similar to matching in many respects, it is not included here because it should be recorded in codes pertaining to analysis or effect size calculation methods.

## 9 Similarity of control group

(1) Internal—Another group from the same pool of Ss—all participants started off as part of one group. (2) External—A group from a patently different pool of participants;. (3) Archival/historical—Data taken from past study (e.g., past experiment; normative data on a test); (4) other; (9) unknown.

**Rationale** (Optional code that is recommended for use in exploration of design effects): All randomized experiments use internal control groups by definition. Some previous meta-analytic evidence suggests that quasi-experiments that use internal controls more closely approximate results from quasi-experiments that use other controls (Shadish & Ragsdale, 1996).

## 10 Activity Level of Control Group

Activity Level of Control Group (1) Passive Control: no treatment, wait list; (2) Active Control: Placebo, treatment as usual; (3) other (9) unknown

**Rationale** (Optional code that is recommended for use in exploration of design effects): This code could be an alternative to code 3 (type of comparison), although it can easily be created from the data in code 3, and code 3 may be preferable to retain more specific information about each study.

## 11 Selection process

(1) Self selection—subjects actively chose their assignment to condition; (2) Other selection #1—someone else selected the units into this condition based on variable(s) related to the outcome (e.g., a physician choosing patients for treatment based on

risk for adverse outcome if untreated); (3) Other selection #2—someone else selected the units into this condition based on variables not obviously related to outcome (e.g., the experimenter picked those subjects who could make certain appointment times). NOTE: Randomized studies should always be coded as other selection #2.

**Rationale** (Optional code that is recommended for use in exploration of design effects): Some previous meta-analytic evidence suggests that quasi-experiments that use other-selection more closely approximate results from randomized studies than quasi-experiments that use self-selection (Heinsman & Shadish, 1996; Shadish & Heinsman, 1997; Shadish & Ragsdale, 1996; Shadish, Matt, Navarro & Phillips, 2000).

## 12 Conditioning of Assignment on Treatment

(1) Conditioned: Which unit received which treatment was a function of some feature of treatment (e.g., volunteers vs. nonvolunteers, completers vs. dropouts, eligibles vs. ineligible); (2) Not conditioned. NOTE: Randomized studies should always be coded as not conditioned.

**Rationale** (Optional code that is recommended for use in exploration of design effects): All randomized experiments by definition assign units to conditions independent of any feature of either the units or the conditions. Nonrandomized experiments that mimic this feature of randomized experiments may produce more accurate estimates of effects.

## 13 Pretest Effect Size

If sufficient data are available, calculate a pretest effect size for whatever comparisons have outcome effect sizes.

**Rationale** (Optional code that is recommended for use in exploration of design effects): Sometimes sufficient pretest statistics exist to compute the same effect size indicator at pretest as is being computed at posttest or follow-up. This pretest effect size has proven to be a very powerful predictor of posttest effect size (Heinsman & Shadish, 1996; Shadish & Heinsman, 1997; Shadish & Ragsdale, 1996). In addition, it has proven useful in exploring the size and nature of selection bias in nonrandomized experiments (Shadish et al., 2000).

## 14 Timing of Measurement in Nonequivalent Groups

(1) Nonequivalent comparison group was measured at the same time as treatment group (i.e., cross-sectional comparison); (2) Nonequivalent comparison group was measured at different time than treatment group (i.e., cohort comparison, historical control).

**Rationale** (Optional code that is recommended for use in exploration of design effects): Some previous meta-analytic evidence suggests that studies using controls measured at a time point different from the treatment group may yield biased effect size estimates (Sacks et al., 1982, 1983).

---

[1] This panel is a subgroup of the Quasi-Experimental Design group, a registered C2 entity.

[2] This is not to preclude the possibility that C2 may eventually develop other databases pertaining to questions about, for example, process and implementation, though it is to presume that such databases would be maintained separately from the database(s) discussed in this policy brief.

[3] **Reviewer Comment:** "I agree with the point of the special significance of randomized trials but, nonetheless, think it is not wise to try to maintain two databases. First, it presumes there is a clear categorical distinction, relating to quality of evidence, between random assignment designs and nonrandom ones. But a random assignment design produces high quality (re internal validity) evidence only with a sufficiently large N (small N randomization doesn't equate groups well), no or trivial attrition prior to outcome measurement, and a clean, relatively unambiguous distinction between the treatment and control conditions. I just reviewed a study with random assignment and subsequent attrition of around 60% in both conditions. Would you put this study in the random assignment database? If not, where would you draw the line on attrition—40%, 20%, 10%? In short, differentiating on the basis of initial randomization by itself doesn't do much in the world of social intervention studies to skim off the high quality designs. If we're going to then consider the other factors relevant to whether it's a "good" randomized design, we'll need consensus criteria on those other factors and the thresholds below which a study fails to make the grade. Developing those will be difficult or impossible and, ultimately, not worth the trouble for purposes of organizing the C2 database of studies. A second, related reason for not maintaining two databases is the implied certification of some studies as categorically better than others that would result. If C2 attains the recognition and stature that we all hope it does, which of the two databases a study is assigned to will be viewed as a kind of quality rating that will be used, and cited, as an expert judgment of the quality of a study's design. This would be OK if it weren't for the issues I raised above. The randomized study database will include some very poor studies that nonetheless used an initial randomization. Thus, while attempting to highlight the special importance of good randomized studies with a separate database, we will also be categorizing some very non-exemplary studies in the gold standard group and implying that, despite whatever other

problems and degradations of the initial randomization they have, they are still categorically better than those in the nonrandomized database. The only way I can see to avoid this implied certification of design quality would be to further differentiate “good” randomized studies from “bad” ones, as discussed above, and this opens up a whole can of worms we’d be wise to avoid.”

[4] Sometimes called cluster randomized experiments.

[5] Reviewer Comment: “I disagree with you here. I would make the distinction based on individual or multiple unit assignment and leave it at that. “Discrepant” units are a statistical analysis flaw that reviewers should pick up later, not folks entering documents into a reference database and “Inadequate” is a judgment call.”

[6] Reviewer Comment: “Should you distinguish: (1) one time series, (2) multiple time series, (3) studies with control series?”.

[7] Reviewer Comment: “In most educational areas that touch on current policy concerns there could be no reviews once you put in the requirement that there be some randomized experiments in the data base. I agree with your proposal, but it has massive implications. One could do some educational reviews, and C2 should do them; but they would be on smaller, tighter topics than what dominates current debates. I wish there was a mechanism in C2 for getting these things done and for putting together a group of folks willing to actually propose and do an experiment. But putting folks from different institutions together requires someone who takes responsibility for the whole and who has the sanctity of the Lord.”

[8] Reviewer Comment: “This is the wrong way round! you won’t know that until the systematic review searching and evaluation processes completed. Make decision to conduct a systematic review because of the importance of the question to be addressed, and only then see what you find”

[9] Reviewer Comment: “I’d prefer wording to the effect that it was generally preferable that there be randomized experiments available. Your stronger wording suggests to me that C2 is not much interested in doing syntheses in areas where randomized designs are not and cannot be done. I acknowledge that point (c) below specifically addresses this case as an exception, but still think the wording above implies that availability of randomized studies is more important than it should be in deciding whether or not to do synthesis on a topic that may be of great significance to practice or policy.”

[10] Reviewer Comment: “I like the position that, in areas where a randomized experiment could be done, if there is none in the data base then one should either not proceed or proceed with the greatest of caution.”

[11] From this code and the three that follow, both total and differential attrition should be

computed.