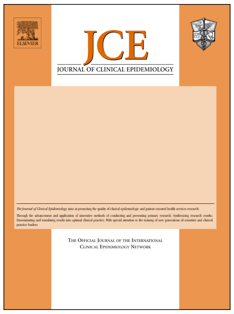

Quasi-experimental study designs series –Paper 9: Collecting Data from Quasi-Experimental Studies

Ariel M. Aloe, Betsy Jane Becker, Maren Duvendack, Jeffrey C. Valentine, Ian Shemilt, Hugh Waddington

PII: S0895-4356(17)30285-8

DOI: 10.1016/j.jclinepi.2017.02.013

Reference: JCE 9350

To appear in: *Journal of Clinical Epidemiology*

Received Date: 22 September 2014

Revised Date: 13 September 2016

Accepted Date: 21 February 2017

Please cite this article as: Aloe AM, Becker BJ, Duvendack M, Valentine JC, Shemilt I, Waddington H, Quasi-experimental study designs series –Paper 9: Collecting Data from Quasi-Experimental Studies, *Journal of Clinical Epidemiology* (2017), doi: 10.1016/j.jclinepi.2017.02.013.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Quasi-experimental study designs series –Paper 9: Collecting Data from Quasi-Experimental Studies

Ariel M. Aloe[1], Betsy Jane Becker[2], Maren Duvendack[3], Jeffrey C. Valentine[4], Ian Shemilt[5], Hugh Waddington[6]

1 University of Iowa, 2 Florida State University, 3 University of East Anglia, 4 University of Louisville, 5 University of Cambridge, 6 International Initiative for Impact Evaluation (3ie)

Acknowledgements

The contributions of Ariel M. Aloe and Betsy J. Becker have been partially supported by grants from the National Science Foundation (NSF DRL- 1550169 and DRL-1252338). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

What is new:

- Methodological differences among QEs must be carefully coded to capture possible sources of variation.
- Information on the statistical and design controls used in QE studies is of critical importance in reviews of QEs.
- If effects do not all estimate the same quantity the meta-analyst should code and examine differences due to effect-size metric.
- When extracting partial (adjusted) effect sizes meta-analysts should document model complexity, and examine its role in between-studies differences.

ABSTRACT

Objective: To identify variables that must be coded when synthesizing primary studies that use quasi-experimental designs.

Study Design and Setting: All quasi-experimental (QE) designs.

Results: When designing a systematic review of QE studies potential sources of heterogeneity – both theory-based and methodological – must be identified. We outline key components of inclusion criteria for syntheses of quasi-experimental studies. We provide recommendations for coding content-relevant and methodological variables, and outlined the distinction between bivariate effect sizes and partial (i.e., adjusted) effect sizes. Designs used and controls employed are viewed as of greatest importance. Potential sources of bias and confounding are also addressed.

Conclusion: Careful consideration must be given to inclusion criteria and the coding of theoretical and methodological variables during the design phase of a synthesis of quasi-experimental studies. The success of the meta-regression analysis relies on the data available to the meta-analyst. Omission of critical moderator variables (i.e., effect modifiers) will undermine the conclusions of a meta-analysis.

KEYWORDS

Meta-analysis, quasi-experiment, bivariate effect size, partial effect size, moderator variables, effect modifiers

In this article we describe the data that should be collected by authors of systematic reviews of quasi-experiments (QEs). Broadly, the categories of data that need to be collected from QE studies are the same as those needed for systematic reviews of experimental studies (i.e., randomized controlled trials), namely information needed to: (a) describe the included studies in terms of their key characteristics, (b) assess the risk of bias in included studies, and (c) compute study-level effect sizes for use in meta-analysis and code predictors for use in moderator analysis. These three categories are not mutually exclusive. Because guidelines for authors on collecting data from experimental studies are well established, we focus here on issues that require particular attention when collecting data from QE studies, referring readers to already published sources of guidance where available (e.g., Higgins et al., 2013; Valentine & Thompson, 2013).

We first consider some general issues in data collection from QEs, then discuss inclusion criteria for systematic reviews of QEs. We discuss recommendations for the coding of content-relevant and methodological variables. After that we describe information needed to compute effect sizes from QE designs. Finally, we provide recommendations for the systematic review of QE studies and highlight areas of future study.

General Issues

When a randomized control trial (RCT) cannot be performed the next best source of evidence is often a well-designed QE (Bärnighausen et al. 2017). Thus, when reviewers want to explore the effects of interventions that have not been studied using RCTs, they may rely on syntheses of QEs (Lavis et al. 2017, Rockers et al. 2017a, Rockers et al. 2017b). The first issue that reviewers will face is what is meant by “QE”. Many different frameworks have been proposed to define and organize QE designs (Campbell & Stanley, 1966; Rockers et al., 2015). For instance, the original work of Campbell and Stanley (1966) labels different designs (e.g., time-series, regression discontinuity) simply as quasi-experimental designs. On the other hand, Rockers et al. (2015) distinguish between different categories of quasi-experiments, based on whether the designs can control for all or only some unobserved confounding. Regardless of the definition of QEs used for a particular systematic review, quasi-experiments are observational studies that attempt to establish causation. Thus, QEs and experiments share the presence of interventions that precede the measurement of the outcome(s). Here we focus on QE designs that include control groups, thus some matters particular to designs such as regression discontinuity and time series (e.g., bandwidth, numbers of observations) are not treated here.

In randomized control experiments, investigators randomly allocated units to intervention and control arms. By definition, QEs lack random assignment of units by investigators. Random assignment creates the expectation of group equivalence on all observed and unobserved variables at baseline. Confounding occurs when individual participant characteristics (e.g., prognostic factors) that are associated with the outcome differ at baseline between participants in treatment and control groups. Researchers can attempt to create a reasonable counterfactual by accounting for confounders using statistical controls or design features such as matching. Sometimes critical covariates can be identified a priori (and thus their presence could be coded), but in the primary studies statistical control variables will be limited to the observed covariates available to the researcher.

Clearly then, information on the statistical controls used in studies will be central in reviews of QEs. These may or may not adequately model the selection process, and therefore may or may not account for a reasonable degree of bias, and the degree to which bias reduction occurs is the key issue for the quality of inference from a QE. Consequently covariates must be examined with care, and one approach to handling the use of covariates is to code those that are significant and function to reduce bias in each primary study. Alternately one could make an overall assessment about whether the collection of controls (statistical or otherwise) appeared to function well for the particular study at hand and reduce bias to some reasonable degree. This more-global kind of judgment is more subjective, and potentially harder to make, but allows for the fact that different covariates may be appropriate for different study configurations.

The potential for heterogeneity in collections of QE studies is greater than in sets of experiments (Deeks et al., 2003). Thus, in the design phase of a systematic review of QE studies potential sources of heterogeneity – both content-relevant (e.g., relating to the content and/or process of the intervention) and methodological – must be identified. Coding these potential sources of heterogeneity allows for empirical assessment of their predictive power. Each systematic review of QE studies will differ in the content-relevant variables (e.g., intervention or population features, etc.) that need to be coded. However, all systematic reviews of QE studies should extract information regarding a core set of methodological variables – and of course the exact methodological issues pertinent to inferences may differ across systematic reviews. Therefore the discussion below should be seen as a starting place for identifying sources of heterogeneity, not as a single set of recommendations for all reviews.

Inclusion Criteria

Inclusion criteria are a central element in any systematic review (Cooper, 2010). Both the Campbell and Cochrane Collaborations ask for clear inclusion criteria in their protocols. The inclusion criteria must be directly connected to the research question(s) and should clearly discriminate between what studies are kept for analysis and what studies are not. For instance, if a systematic review of QE studies is intended to include only QEs which used a regression discontinuity approach, the inclusion criteria should clearly state this.

A key element of the inclusion criteria for a review of QEs is information about how control variables will be treated. Because selection bias is a major concern, the reviewer must have a deep understanding of how selection effects operate in the studies in the review, then use this knowledge to outline potential important control variables (e.g., age, gender, etc.). The reviewer also must decide how the included studies *should* have handled those variables. For example, inclusion criteria may allow studies to report on participants of any age. For age to be investigated as a possible source of heterogeneity across studies (i.e., to be able to use age as a moderator), it must be coded at the study level, and it must vary across studies. However, just coding the average age of participants does not achieve control of age-related selection bias within each study. Thus the inclusion criteria might also specify that age should be controlled within each study, for example, by using age as a covariate or a blocking variable. In such a case the reviewer would not select a study with a broad age range unless age was used as a covariate. On the other hand, a study with a sample of only elderly participants in a narrow age range may be eligible even if it does not covary or match on age, simple because participant ages do not vary greatly, thus care controlled by design (i.e., by restricting the sample).

Coding Content-relevant and Methodological Variables

The particular variables that are theoretically relevant for any given review will depend on the topic of interest. Thus, choosing specific content-relevant variables to extract must be done in the context of the particular topic under consideration in the systematic review. This is true for all research syntheses, not just those that include QE studies. Nevertheless, some key variables can be identified. Because QEs typically evaluate interventions, aspects of the treatment are critical. Duration, intensity, and information about implementation are needed. If a variety of individuals implement the treatment, their characteristics should be tracked. Variables such as amount of training received or experience of the implementers may affect strength of the treatment effect. Information on fidelity of intervention will enable to reviewers to examine whether the strength of an intervention varies according to how well it is put in place. Likewise, whether a treatment has a manual, and whether the manuals are used in practice, may relate to program efficacy. Population target, types of outcomes and how they were measured, and other contextual factors relevant for the intervention should be extracted as well. Perhaps obviously, the meta-analyst should extract setting characteristics as well. Frameworks such as PICOS (Patients, Intervention, Comparisons, Outcomes, Study Design; Richardson et al., 1995) or MUTOS (Methods, Units, Treatments, Observations, Settings; Cronbach, 1982; Becker, 1996; Becker & Aloe, 2008) can guide selection of relevant features.

Careful attention also needs to be paid to the designs and analyses used in the primary studies. If these are dissimilar, inclusion of a meta-analysis (i.e., quantitative summary) in the systematic review must be justified. Thus, reviewers should not only code how the primary study analyzed the data, they should also code the specific predictors used, and details of how the outcome(s) were measured. As mentioned above, information should be coded on whether the important statistical assumptions of the QE analysis were met. The success of adjustments for confounders and selection bias in a QE relies on meeting the assumptions of the analyses performed.

In the same manner, when the QE design includes a control group the presence (or lack) of evidence of the comparability of groups on key variables should be coded. At minimum, the presence or absence of information on assumptions (e.g., in the form of dummy variables) and baseline group comparisons should be coded. Then methods such as meta-regression can be used to explore systematic differences between primary studies that do or do not provide this information.

For a QE study using a regression discontinuity design (Thistlewaite & Campbell, 1960) the reviewer should code information about the assumption that the treatment effect is discontinuous at the cutoff. When specific information about assumptions is not given, reviewers should request such information from the primary study authors. We see this as being analogous to the queries meta-analysts make of RCT designs to verify that randomization was successfully implemented, that attrition was low, and so forth.

Coding Information on Effect Sizes

The term effect size is commonly used in meta-analysis to refer to the study outcome. In the social sciences, some meta-analysts refer to three families of effect sizes, namely the *d* family, *r* family, and *p* family (Lipsey & Wilson, 2001). The *d* family includes standardized mean differences (for a continuous outcome and a grouping variable). The *r* family includes correlation coefficients (usually between two continuous variables). The *p* family includes proportions or quantities related to proportions such as odds ratios (reflecting relations between dichotomous variables). Effect sizes can often be transformed from one metric to another (see Lipsey & Wilson, 2001). The choice of what quantity best represents study outcomes will depend on the design and analysis in the primary study, and the desired inferences for the systematic review. All three families may be appropriate as effects, depending on the QR design.

Bivariate and Partial Effects

The distinction between bivariate effect sizes and partial (i.e., adjusted) effect sizes has until recently been largely ignored (some notable exceptions are Aloe, 2014; Aloe & Becker, 2012; Aloe & Thompson, 2013; Keef & Roberts, 2004). However, given the importance of control variables in QEs, this distinction is highly relevant for effect-size computation for QEs. Bivariate effect sizes in all effect-size families portray relationships between two variables. In contrast, partial effect sizes gauge the strength of relationships between two variables adjusting for the effects of other variables in the model. For instance, one primary study may report the relationship between breastfeeding and the risk of infant overweight as an odds ratio – a bivariate effect. Another study may report the relationship between breastfeeding and the risk of infant overweight as an odds ratio adjusted by (or conditional on) gender, birth weight, and socioeconomic status, which would be a partial effect. This distinction between bivariate and adjusted effects is important because the magnitudes and even the directions of adjusted effects can differ greatly from those of unadjusted effects. Also certain control factors may lead to large adjustments whereas others have no impact. Thus coding the presence of key control variables allows the reviewer to analyze variation among partial effects if a meta-analysis is conducted.

Because QE designs aim to assess the effectiveness of treatments, the fundamental effect of interest is likely to be based on mean differences or comparisons of counts or odds. Because of the complexities of QEs, the effect size will nearly always be something other than the simple standardized mean difference (*d*, as in Hedges & Olkin, 1985), or the simple odds ratio. Most effect sizes will likely be partial effects and their magnitudes will depend on what is adjusted for in each study. Therefore meta-analysts must code detailed information about the specific controls used – either design controls, statistical control variables, or both – during data extraction.

For each study (or sample, for studies with multiple samples), the meta-analyst should extract both an estimate of effect size and its standard error (SE). If a study does not report standard errors or the information needed to derive them – for example, if a study reports an odds ratio but does not provide the underlying success rates – the meta-analyst may be able to compute them from related test statistics (e.g., *t* tests of slopes). Effect estimates can also be obtained from observed significance levels, but if the SE cannot be obtained it may have to be imputed, or the study may need to be dealt with in a narrative fashion.

Example. In this section we illustrate the differences between bivariate and partial effect sizes, for the case of standardized mean differences. When a randomized control trial is conducted and sufficient statistics reported, it is straightforward to compute the standardized mean difference. Assume that the study compares a treatment group with a control group. The standardized mean difference is defined (Hedges & Olkin, 1985) as

$$g = \frac{\bar{Y}^T - \bar{Y}^C}{s}, \qquad (1)$$

where $\bar{Y}^T$and $\bar{Y}^C$are the sample means of the treatment and control groups, respectively, and *s* is the pooled sample standard deviation. When means or standard deviations are not provided, meta-analysts typically obtain *g* by transforming other reported statistics (see Lipsey & Wilson, 2001, for specific formulas). For instance, if a *t* test between independent means is conducted and the *t* statistic (*t*) is reported, then a little algebra yields

$$g = t\sqrt{\frac{n_T + n_C}{n_T n_C}}, \qquad (2)$$

where n_T and n_C are the sample means of the treatment and control group. Equations 1 and 2 are algebraically equivalent.

Because QEs tend to rely on regression analysis, the outcome of interest is predicted by the focal predictor (i.e., the variable representing group membership) and other observed covariates (i.e., the analysis of covariance or ANCOVA model). Keef and Roberts (2004) proposed a partial effect size for group differences based on ANCOVA. Specifically, their model was

$$Y_j = \alpha + \gamma D_j + \beta_2 X_{2j} + \ldots \beta_p X_{pj} + e_{j,}$$

where Y is an outcome score for person j, D is a dummy variable representing a treatment or group effect, and X_2 through X_p are covariates. The errors e_j are assumed to have common variance σ_e^2 estimated as s_e^2. Keef and Roberts proposed using $g_p = \hat{\gamma} / s_e$ as a partial index of treatment effects, since $\hat{\gamma}$ represents an adjusted mean difference (accounting for all covariates in the model) and s_e^2 is the residual variance – essentially the variance of the Y scores, adjusting for the effects of all predictors. If the meta-analyst does not have access to s_e the standard deviation (s_Y) of the outcome can be used. Moreover, for some studies meta-analysts may prefer to standardize the partial effect size by the within groups pooled standard deviation to avoid extra artificial variability in s_e^2 (Johnson & Eagly, 2000). Other partial effect sizes estimated as a function of t-test statistics, such as partial correlations, can overestimate the magnitude of partial effects. This occurs because typically covariates reduce the size of the standard error of the slope of interest (i.e., the denominator of the test) relative to those in other studies.

In ANCOVA scenarios with significant covariates, the bivariate and partial effects will typically disagree; thus, $g \neq g_p$. If the primary study achieves the goal of adjusting for all of the 'right' variables, this partial effect will approximate the bivariate effect obtained from a randomized control trial (i.e., $g \cong g_p$). However, adjusting statistically many not be the same as controlling by design. Evidence on when such effects may be comparable comes from a study conducted by Shadish and colleagues (2004). Students were randomly assigned to two assignment mechanisms (i.e., random assignment and nonrandom assignment) (Shadish, Clark, & Steiner, 2004) then were assigned to treatment and control conditions using that mechanism. Data from the randomized and nonrandomized groups were analyzed and the extent to which the nonrandomized groups results deviated from the experimental results was conceived of as bias. The authors reported that both linear-regression and propensity-score adjustments reduced 95% and 96% of this observed bias, respectively. However, typically, a great deal must be known about the topic of interest and the mechanisms behind how the treatment works for adjustments to be this effective. For example, Shadish et al. first explored theoretical considerations related to why participants might choose a particular intervention condition, and then measured several indicators of ability and other individual differences, demographics, and attitudes toward the intervention; these were used as control variables.

Primary Studies Reporting Multilevel Models

Clusters of individuals in neighborhoods, villages, or other units such as medical practices may be recipients of treatments studied in QEs. The nesting involved in clustered study designs needs to be considered in the computation and analysis of effect sizes (e.g., Hedges, 2007; 2011). Reviewers should extract the intraclass correlation (or information needed to compute it), and the number of units that are represented at different levels (e.g., patient at level

1, clinic at level 2), from each multilevel model. This information can then be used to correct the effect size and its variance (see Hedges, 2007).

Dependence

One important assumption of typical meta-analytical techniques is independence of the effect sizes. When a study reports results for multiple outcomes, or if the same outcome is reported at multiple time points, the several effects will be interdependent. Or, sometimes primary studies will report several models, often to show the contributions of different sets of predictors. Some meta-analysts extract and analyze effects from every model in each primary study. However if all such effects are included in one meta-analysis, the independence assumption fundamental to univariate meta-analytic statistics is violated. In addition studies reporting many models will have undue influence if the effects from them are treated as independent. To account for such dependence requires knowing or imputing information on correlations among variables (which should be recorded if they are reported). A range of approaches exist to modeling dependence (see Becker et al., this issue; Gleser & Olkin, 2009) but these can be very hard to implement because data on the degree of dependence is often unavailable. A simple approach when several competing models are reported in a primary study would be to select the one that best controlled for potential biases, and to extract the effect-size index from that analysis.

Effect-Size Metric

Given the different designs and analyses used and reported in primary studies, the metrics of the available estimated effect sizes may not always be immediately comparable. A systematic review of QEs must be particularly careful about this, because of the myriad of different designs and analyses used within QEs. For instance, assume that a reviewer is interested in the effect of maternal alcohol consumption on birth weight. It is conceivable that one QE might use an instrumental-variables approach (e.g., state alcohol sales tax as an instrumental variable) while another QE uses propensity score matching (e.g., participants matched on their state's alcohol sales tax and ten other variables). Several questions need consideration: Are effect sizes from these different QE designs comparable? Should effect sizes from these two studies be combined into the same aggregate? If they are estimating clearly different quantities, the answer is no. Studies with different designs may need to be analyzed separately, especially when design type is confounded with things such as the population type or the type of effect size that is reported (as in Kownacki & Shadish, 1999). Moreover, some QE studies may report the average treatment effect on the treated (ATT), whereas other QE studies investigating the same matter may report the average treatment effect (ATE), which involves both treatment and control groups (Holland, 1986; Imai et al., 2008). Yet another QE may report the local average treatment effect (LATE). Are these three quantities comparable? Should these three quantities be combined into the same aggregate? The answers to these questions will depend on the particularities of the primary studies, though users should be prepared for the answer to be "No". That said, the impact of combining these three quantities in meta-analysis is an area that deserves further investigation. Regardless of whether different estimators are thought to be roughly comparable, data on the nature of the specific QE designs and analyses must be extracted from studies. That information can be used to empirically investigate differences in estimated effects and variability within and between specific designs (see Becker et al., same issue). Arguably what should be combined in a meta-analysis depends on the specific research question. In some cases differences in type of effect size can be seen as potential sources of variability. However, in other situations where quite disparate quantities are being estimated, having such differences would be problematic for the quantitative synthesis.

General Practical Recommendations

As is true for primary studies, decisions made during each step of a systematic review have consequences for the next steps of the review. Thus, careful planning and consideration of a protocol is a vital component of any systematic review. For instance, if an important rule is omitted from the inclusion criteria, the collection of studies may not accurately represent the desired population (e.g., if a systematic review was not specific about the operational definition of QE, relevant studies may be omitted and/or irrelevant studies may be included). If a particular factor is ignored at the coding stage, information will not be available for the reviewer in the data-analysis stage (e.g., if a systematic review did not code which covariates were used in the QEs, this information will not be available for use as potential moderator variables). Thus, the design phase (i.e., protocol development) of any systematic review of QEs is critical. Reviewers should anticipate, then code, all possible factors (substantive/theoretical, design, analysis, and so on) that may lead to differences in QE outcomes.

Many claims about similarities and differences among diverse QE designs and analyses can be empirically explored when covariates are coded. When the coded covariates are too numerous to be included individually in a single analysis the meta-analyst may be able to create higher level constructs that represent several related covariates. Similarly, different QE designs adjust for confounders and selection bias through different mechanisms (see for example Hernan, Hernandez-Diaz, & Robins, 2004). By coding the specifics regarding designs and analyses of each included primary study the relationships of these features to the treatment effects can be explored as well.

Conclusions

The data-collection stage of any meta-analysis is crucial because other stages such as data analysis will depend on what has been coded. In addition to effect-size information, study characteristics that could potentially modify the magnitude of effects across studies should be systematically coded.

Although many alternative QE designs exist, the overall goal of all of the designs is to control for endogenous selection in the assignment of people to intervention and control groups, similar to trials. When primary studies report on QEs, they commonly use analyses based on regression models; the predictors included in these models are key to ensuring that the studies provide credible evidence about treatment effectiveness. However, no statistical technique can correct for poor design and/or poor data in a primary study or a meta-analysis. Thus, careful attention must be given to the collection of data from QE studies included in a systematic review.

We have provided suggestions about what variables should be coded when synthesizing QEs, and outlined some considerations regarding effect sizes. As is true for any meta-analysis of aggregate data, the goal in the data-extraction phase in a synthesis of QEs is to extract from each primary study the best evidence on the effects of the intervention. Attending to those features that ensure group comparability in comparative QE designs should support credible analyses and synthesis conclusions.

References

Aloe AM, Becker BJ. An effect size for regression predictors in meta-analysis. **J Educ Behav Stat**, 2012; 37: 278-297.

Aloe AM, Thompson CG. The synthesis of partial effect sizes. **J Soc Social Work Res**, 2013; 4: 390-405.

Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. **J Am Stat Assoc**, 1996; 91: 444–55.

Bärnighausen T, Tugwell P, Røttingen J, Shemilt I, Rockers P, Geldsetzer P, Lavis J, Grimshaw J, Daniels K, Brown A, Bor J, Tanner J, Rashidian A, Barreto M, Vollmer S, Atun R. Quasi-experimental study designs for evaluating practice, programs and policies: uses and value. **Journal of Clinical Epidemiology** (this issue), 2017.

Becker, BJ. The generalizability of empirical research results. In: Benbow CP, Lubinski D, eds. **Intellectual Talent: Psychometric and Social Issues** (pp. 362-383). Baltimore: Johns Hopkins Press, 1996.

Becker B, Aloe AM, Duvendack M, Stanley TD, Valentine JC, Fretheim A, Tugwell P (2017). Quasi-experimental study designs for evaluating practice, programs and policies: Synthesizing evidence for effects collected from quasi-experimental studies presents surmountable challenges. **Journal of Clinical epidemiology** (this issue).

Becker BJ, Aloe AM. **A framework for generalization in meta-analysis: Medical and social-science examples**. Invited presentation at the 16th Merck-Temple Conference on Biostatistics, Philadelphia, PA, 2008.

Campbell DT, Stanley JC. **Experimental and Quasi-experimental Designs for Research**. Rand McNally: Chicago, Illinois, 1966.

Cooper H. **Research Synthesis and Meta-analysis: A Step-by-step Approach** (4th Ed.). Thousand Oaks, CA: Sage, 2010.

Cronbach LJ. **Designing Evaluations of Educational and Social Programs**. San Francisco: Jossey-Bass, 1982.

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomized intervention studies. **Health Technol Assess**, 2003, 7: 1–179.

Gleser LJ, Olkin I. Stochastically dependent effect sizes. In Cooper H, Hedges LV, Valentine JC, eds, **The handbook of research synthesis and meta-analysis** (2nd ed.) (pp. 357–376). New York: Russell Sage Foundation, 2009.

Hedges LV. Effect sizes in cluster-randomized designs. **J Educ Behav Stat**, 2007; 32: 341-370.

Hedges, LV. Effect sizes in three level designs. **J Educ Behav Stat**, 2011; 36: 346-80.

Hedges LV, Olkin I. (1985). **Statistical Methods for Meta-analysis**. Orlando, FL: Academic Press.

Higgins JPT, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC, Tugwell P, Wells G. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. **Res Syn Meth**, 2012; 4(1): 12-25.

Johnson BT, & Eagly AH. (2000). Quantitative synthesis of social psychological research. In Reis HT Judd CM (Eds.), **Handbook of research methods in social and personality psychology** (pp. 496-528). New York: Cambridge University Press.

Keef SP, Roberts LA. The meta-analysis of partial effect sizes. **Brit J Math Stat Psych** 2004*;* 57(1): 97-129.

Kownacki RJ., Shadish WR. Does Alcoholics Anonymous work? The results from a meta-analysis of controlled experiments. **Subs Use Misuse** 1999; 34(13):1897-916.

Lavis J, Bärnighausen T, El-Jardali F. Quasi-experimental study designs for evaluating practice, programs and policies: Supporting the production and use of health systems research syntheses that draw on quasi-experimental study designs. **Journal of Clinical Epidemiology** (this issue), 2017.

Richardson WS, Wilson MC, Nishikawa J, Hayward RSA. The well-built clinical question: A key to evidence-based decisions. **ACP J Club**, 1995; 123(A12-A13).

Rockers PC, Røttingen, J-A, Shemilt I, Tugwell, P, Bärnighausen T. Inclusion of quasi-experimental studies in systematic reviews of health systems research. **Health Policy**, 2015; 119(4): 511-521.

Rockers PC, Tugwell P, Grimshaw J, Oliver S, Atun R, Røttingen J, Fretheim A, Ranson MK, Daniels K, Luiza VL, Bärnighausen T. Quasi-experimental study designs for evaluating practice, programs and policies: Strengthening global capacity for evidence synthesis of quasi-experimental health systems research. **Journal of Clinical Epidemiology** (this issue), 2017.

Rockers PC, Tugwell P, Røttingen JA, Bärnighausen T. Quasi-Experimental Study Designs for Evaluating Practice, Programs and Policies: Realizing the Full Potential of Quasi-Experiments for Health Research. **Journal of Clinical Epidemiology** (this issue), 2017.

Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing. **J Am Stat Assoc**, 2004; 103: 1334-1344.

Thistlewaite D,Campbell D. Regression-Discontinuity Analysis: An alternative to the ex post facto experiment. **J Educ Psych** 1960; 51: 309–317

Valentine JC, Thompson SG. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. **Res Syn Meth**, 2013; 4(1): 26-35.