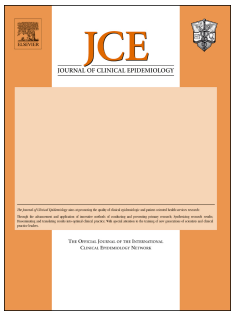

Quasi-experimental study designs series – Paper 10: Synthesizing evidence for effects collected from quasi-experimental studies presents surmountable challenges

Betsy Jane Becker, Ariel M. Aloe, Maren Duvendack, T.D. Stanley, Jeffrey C. Valentine, Atle Fretheim, Peter Tugwell

PII: S0895-4356(17)30286-X

DOI: 10.1016/j.jclinepi.2017.02.014

Reference: JCE 9351

To appear in: *Journal of Clinical Epidemiology*

Received Date: 22 September 2014

Revised Date: 22 February 2017

Accepted Date: 22 February 2017

Please cite this article as: Becker BJ, Aloe AM, Duvendack M, Stanley TD, Valentine JC, Fretheim A, Tugwell P, Quasi-experimental study designs series – Paper 10: Synthesizing evidence for effects collected from quasi-experimental studies presents surmountable challenges, *Journal of Clinical Epidemiology* (2017), doi: 10.1016/j.jclinepi.2017.02.014.

Quasi-experimental study designs series – Paper 10:

Synthesizing evidence for effects collected from quasi-experimental studies presents surmountable challenges

Betsy Jane Becker[1], Ariel M. Aloe[2], Maren Duvendack[3], T. D. Stanley[4], Jeffrey C. Valentine[5], Atle Fretheim[6], Peter Tugwell[7]

1 Florida State University, 2 University of Iowa, 3 University of East Anglia, 4 Hendrix College, 5 University of Louisville, 6 Norwegian Knowledge Centre, 7 University of Ottawa

DRAFT: Not for citation.

Synthesizing evidence from quasi-experimental studies presents surmountable challenges

ABSTRACT

Objective: To outline issues of importance to analytic approaches to the synthesis of quasi-experiments (QEs), and to provide a statistical model for use in analysis.

Study Design and Setting: We drew on the literatures of statistics, epidemiology, and social-science methodology to outline methods for synthesis of QE studies. The design and conduct of quasi-experiments, effect sizes from QEs, and moderator variables for the analysis of those effect sizes were discussed.

Results: Biases, confounding, design complexities and comparisons across designs offer serious challenges to syntheses of QEs. Key components of meta-analyses of QEs were identified, including the aspects of QE study design to be coded and analyzed. Of utmost importance are the design and statistical controls implemented in the QEs. Such controls and any potential sources of bias and confounding must be modeled in analyses, along with aspects of the interventions and populations studied. Because of such controls, effect sizes from QEs are more complex than those from randomized experiments. A statistical meta-regression model that incorporates important features of the QEs under review was presented.

Conclusion: Meta-analyses of quasi-experiments provide particular challenges, but thorough coding of intervention characteristics and study methods, along with careful analysis, should allow for sound inferences.

KEYWORDS

Meta-analysis, quasi-experiment, effect size, risk-of-bias, moderator variables, confounding

WHAT-IS-NEW BOX

Meta-analyses of quasi-experiments must investigate moderators that capture key features of the interventions examined and methods used in the primary studies.
The use of statistical and design controls in quasi-experiments leads to complexities in representing QE study effects, as well as in analysis of those effects.
Exploring potential sources of bias and confounding is especially critical when modeling effects from quasi-experimental designs. The use of meta-regression models facilitates such analyses.

1 INTRODUCTION

Syntheses that include quasi-experiments must consider a variety of design and analysis issues that greatly increase the complexity of the meta-analysis process. The growing importance of synthesizing quasi-experiments is evidenced by a recent special issue of *Research Synthesis Methods* (Volume 4, Issue 1), though research on this topic dates back 30 years (Bryant & Wortman, 1984; Wortman, 1992). In this work we discuss key elements of potential analytic approaches to the synthesis of quasi-experiments, and provide a broad statistical model for use in analysis.

Other papers in this themed issue of the *Journal of Clinical Epidemiology* describe how quasi-experimental studies can be identified for evidence synthesis (Glanville et al. 2017), how data is best collected from quasi-experimental studies (Aloe et al. 2017), and how the global capacity for including quasi-experimental studies in evidence synthesis can best be expanded (Lavis et al. 2017, Rockers et al. 2017). In this paper, we begin with a brief discussion of several definitions of quasi-experiments (QEs) and describing some challenges that arise in synthesizing QEs. We next describe the information required to carry out such a synthesis. This includes information on effect sizes, study features, and the details of the models examined in the primary QE research. We conclude with potential unresolved issues in this domain.

2 QUASI-EXPERIMENTAL DESIGNS

Campbell and Stanley (1966) introduced the term "quasi-experiment" in their seminal book on design of studies. They described QEs as "settings in which the research person can introduce something like experimental design… even though he lacks the full control over the scheduling of experimental stimuli… which makes a true experiment possible" (p. 34). They outlined ten different QE designs. Rockers et al. (2015) drew on this definition and several others to arrive at a different definition, writing that QEs "… estimate causal effect sizes using exogenous variation in the exposure of interest, which is not directly controlled by the researcher" (p. 511). They continue "… five commonly used designs … fit our definition of quasi-experiments: natural experiments, instrumental variable analyses, regression discontinuity analyses, interrupted times series studies, and difference studies including controlled before-and-after designs, difference-in-difference designs and fixed effects analyses of panel data" (p. 511). Rockers et al. distinguish between study designs that control all confounding (observed and

unobserved), observed and some unobserved confounding, or only observed confounding. These distinctions emphasize the important role of confounding in the synthesis of QEs. Bärnighausen et al. (2017) describe the assumptions that need to be met in different types of quasi-experimental studies that replace the (strong) unconfoundedness assumption in non-experimental studies.

Various other terms are used across different disciplines to refer to the diverse array of nonrandomised studies (e.g., observational studies, natural experiments, cohort designs) not all of which are considered QEs. For example, the Cochrane Collaboration Handbook (section 13.2.2 at http://handbook.cochrane.org/chapter_13/13_2_2_guidance_and_resources_available_to_support_review.htm) tables the features present in 18 non-randomized study designs, and urges reviewers to use those design features to determine which studies might be included in a review. Also Wells et al. (2013) provide a checklist of design features that may help classify a study design. For the meta-analyst, relying on the features of studies may be more informative than deciding on inclusion based on global labels (with no discussion of details of design).

These lists, along with Campbell and Stanley's ten original designs from 1966, make clear that QEs are complex, are not themselves internally coherent, and are labelled in diverse ways across fields, which leads to some of the myriad challenges encountered in syntheses of QEs. These issues are compounded in reviews that combine QEs with other kinds of study designs, most critically with experiments.

A fundamental feature of experiments is that units of interest are assigned to treatment conditions randomly and independently (or randomly with explicit constraints, as in randomized blocks designs). In QEs, the treatment or exposure is not so tightly controlled, and randomization is not, or cannot be, fully achieved. Because treatments are not randomly assigned to units (or vice versa), QEs usually involve designs and analyses that attempt to control confounders and other biases in other ways. However, the fact that an analysis includes control variables does not make it a QE. Rockers et al. (2015) and Bärnighausen et al. (2017) further elaborate on confounding control in quasi-experimental studies.

Wells et al. (2013) provide a checklist of other issues for reviewers to consider when attempting a review that may include QEs (or other nonrandomized studies, in their

nomenclature). They describe protocol development for the review, primary-study assessment, and outcome assessment (within study). We draw on their checklist because it is quite thorough, but we do not deal with protocol development in our discussions, because in syntheses we deal only with completed studies. (Reeves et al. (2017) provide an update and extension of this checklist in this issue.)

QEs can vary widely in terms of how control is exercised, and how potential confounding is handled. This diversity leads to design-based and/or analysis-based variation in study outcomes. This occurs both across designs (e.g., comparing natural experiments versus time-series designs) and within any given type of design. Design decisions such as selecting restricted samples may be made to "control" variance that in other studies is dealt with statistically (e.g., with covariates or stratification). For example, age is often a relevant control variable in health studies. However, some studies may control for age by restricting the sample to a particular age group while other studies include a wider variety of participant ages and use age as a covariate in the analytic model. Thus control variables may be part of the analysis in some QEs, whereas they will not in others. In both cases, control is exerted. Other design and analysis approaches may be used to deal with potential confounding and other biases. Wells et al. (2013) give extensive suggestions on how to evaluate whether primary studies have dealt with confounding. Extracting (Aloe et al., this issue) and analyzing information on biases and controls is critical in a synthesis of QEs. Understanding the relevant counterfactuals for included designs may be helpful in choosing what features to extract.

In addition, design features and other study characteristics such as the nature of the population or features of the interventions or outcome measure used may be confounded across studies in any review of research, leading to entangled conclusions about effects (Lipsey, 2003). This will likely be an issue when different types of QEs are summarized, especially if quasi-experiments and true experiments are synthesized together. As one example, Kownacki and Shadish (1999) summarized studies of a variety of Alcoholics Anonymous (AA) rehabilitation programs. They discovered that randomized controlled trials (RCTs), in the main, had examined populations of persons mandated to attend AA because of drunk driving or other offenses. These randomized studies showed poorer results for AA than did nonrandomized studies. In contrast, the nonrandomized designs (some of which appeared badly biased) examined other types of attendees.

Kownacki and Shadish found that study design and subpopulations studied were greatly confounded, with biases in the mix as well. The feature confounded with study design was subpopulation, but any study feature could be confounded with study design (e.g., QEs may aim at larger samples of the population than RCTs, thus treatments may be less well implemented because of sheer study size). If QEs and RCTs are included in one synthesis, their features must be coded and examined statistically. Graphical displays such as grouped forest plots will also prove useful. This holds as well for syntheses with different kinds of QEs in the mix (and no RCTs). Aloe and co-authors (this volume) discuss the coding of QE studies in detail.

Because study features may be confounded with each other and with aspects of design in particular, meta-analysts should always examine the correlations among study features of interest. High correlations among study features mean that clear conclusions may not be reachable for a particular research domain (e.g., Becker, 1986, pp. 203-204). This also means that multiple predictor variables may provide competing explanations of the variation in study results. If correlated predictors appear together in a meta-regression model, the issue of multicollinearity may also arise. Rubin (1992) noted that all meta-analyses are at high risk of confounding of study features, due to the survey-like nature of the data-collection process.

To summarize, the large heterogeneity inherent in QEs and other study designs presents both challenges and opportunities to reviewers. However, as Berlin (1995) noted in an assessment of the potential to synthesize observational studies, “heterogeneity is our friend.” When diverse studies are analysed properly, heterogeneity can lead to better understandings of phenomena of importance. Take the AA example above: Coding and analyzing differences due to each study’s research design, population characteristics (including whether participation was mandatory), and their interactions will tease out much more useful information about the effectiveness of AA than any simple meta-analysis of just the subset of RCTs or QEs involved.

3 WHAT IS NEEDED TO DO A META-ANALYSIS OF QES?

3.1 Effect sizes

Meta-analyses require measures of effect magnitude that can be compared across studies. Effect sizes for QEs are discussed in detail by Aloe et al. (this issue). 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 they may be available from related test statistics (e.g., *t* tests, *p* values). Effect estimates can be obtained from observed significance levels, but if the SE cannot be obtained or imputed the study may need to be dealt with in a narrative fashion.

3.1.1 Partial and bivariate effects. Because QEs aim to assess the effectiveness of treatments, the fundamental effect of interest is likely to be based on a mean difference or comparison of counts or odds. However, because of the complexities of quasi-experimental designs, the effect size will nearly always be something other than a simple standardized mean difference (*d*, as in Hedges & Olkin, 1985), or a simple odds ratio. Aloe et al. (this issue) discuss the computation of effect sizes for QE studies. Thus here we only note that when varied analytical approaches are used in QEs, their effect sizes will likely not be estimating "the same" (i.e., mathematically identical) parameters across studies. Most effect sizes will likely be partial (adjusted) effects, arising for instance from multiple regression analyses where covariates control for confounds and explain variation in the outcome (e.g., Aloe & Becker, 2011; Keef & Roberts, 2004). Their magnitudes will depend on what is included in each study's analysis. Therefore during data extraction meta-analysts must code detailed information about the covariates and design approaches used – be they design controls, statistical control variables, other predictors of interest, or all of the above.

If some studies in a synthesis report bivariate effects and others report partial effect sizes, the reviewer must decide how to proceed. One option is to include all effects and ignore differences between them. This is potentially problematic because partial effect sizes estimate different parameters and can be larger, smaller, or even opposite in sign from bivariate effects because of the variables that were adjusted for (Aloe, 2014). We recommend that reviewers record whether each effect is bivariate or partial (and also extract information on use of specific control variables in each QE), and test for differences in effect sizes across effect-size types. Distinctions can also be made between QE studies that control for the effects of other variables by design (and may provide bivariate effect sizes) versus other QEs with partial effects that adjust for other variables statistically.

Several analytic options are possible. Most simply, the meta-analyst could report separate analyses for bivariate effects and partial effects. This strategy also may be valuable when QEs and experiments both appear in a synthesis. Alternately one could summarize all bivariate and

partial effects together, using a meta-regression model to capture what is controlled in each study.

3.1.2 Effects from very different designs. Studies with quite different designs may need to be analyzed separately. Aloe et al. (this issue) pointed out that even within different types of QE designs primary studies could be estimating different types of study-level effects. Separating studies by design may be especially useful when design type is confounded with the type of effect size that is reported or with population type, as in Kownacki and Shadish (1999). However doing nothing more than just separating the sets of studies misses an important opportunity to statistically control for study-level quality differences, biases, and confounders. Thus the meta-analyst needs to consider whether the parameters being estimated in the array of studies collected in the review are commensurable "enough", and thus could be analyzed together, or whether they are so fundamentally different estimators thus should be kept apart.

3.1.3 Multiple effects. Sometimes primary studies report several models – either examining distinct subsamples of participants or showing contributions of different subsets of predictors. If a study reports models for non-overlapping, independent subsamples, effects can be extracted from each without great concern for statistical dependence.

In contrast, when several models are estimated for the same participants it can be problematic to include the multiple effects that represent them if they are treated as independent effects. For example, a study may examine how a treatment impacts two outcomes, say blood pressure and quality of life. Or researchers using regression methods to estimate the impact of an intervention on one outcome may present findings from models where they have adjusted for many, few, or no covariates, and the results may differ from one model to another. Some meta-analysts have extracted effects from all models given in each primary study, even when they are estimated for a single sample. This practice leads to violations of the independence assumption required by univariate analyses of effect sizes (Becker, Aloe, & Olkin, under review), and privileges results from studies that report more models by giving them more weight in the analysis.

If a study has examined how a predictor (e.g., an intervention) relates to two different measures of the same outcome, a variety of different approaches exist for handling the well-understood covariation between the effect sizes for those two relationships. Reviewers can use either some *a priori* objective criterion for selecting one estimate per study (e.g., taking an

average of the two effects, selecting the estimate that maximizes similarity with other studies in the meta-analysis, randomly selecting one estimate) or employ more sophisticated analyses that accommodate within-study dependence. Becker (2000) and others have discussed at length potential choices of a single effect. Generalized least squares (GLS; Berkey et al., 1996; Raudenbush, Becker & Kalaian, 1988; among others) and computing cluster-robust standard errors (Hedges, Tipton, & Johnson, 2010) adjust the variance-covariance matrix for within-study dependence and thereby correct confidence intervals and resulting inferences. Unbalanced panel, multi-level, and hierarchical linear (HLM) models have been widely used by meta-analysts to accommodate within study dependence (Kalaian & Raudenbush, 1996; Rosenberger & Loomis, 2000; Stanley & Doucouliagos, 2012; Van Den Noortgate et al., 2013). Some approaches described in the literature are simple, but are not effective at accounting for dependence (e.g., including effects for all outcomes or models, and weighting them by the inverse of the number of reported estimates per study).

Criteria for selecting a single model to represent the meta-analysis parallel the suggestions for addressing situations in which a study presents more than one measure of the same outcome. For example, researchers could select one model based on objective criteria stated *a priori*, such as the model with the most predictors, the model that best approximates the meta-analyst's view of the selection process, or the model that includes a specific pattern of covariates. Meta-analysts might also consider requesting that the primary-study authors run a specific model (i.e., a model with the meta-analysts' preferred set of covariates), or could request the original data run the desired analysis themselves.

3.2 Key theoretical variables

Perhaps obviously, the meta-analyst should extract critical population and setting characteristics. Frameworks such as PICOS (Patients, Intervention, Comparisons, Outcomes, Study Design; Richardson et al., 1995) or MUTOS (Methods, Units, Treatments, Observations, Settings) can guide selection of relevant features (Cronbach, 1982; see also Becker, 1996; Becker & Aloe, 2008). It may be useful to define an ideal target study (Sterne et al., 2013) and characterize studies based on its critical features. Slavin argued that "…a useful organizing principle is the need to be strict on issues with potential for bias and liberal on issues that have little such potential" (2008, p. 7). Other between-studies differences that might affect reported study outcomes should also be extracted.

3.3 Design information

3.3.1 Nature of relevant comparison. Wells et al. (2013) argued that the first defining characteristic of a nonrandomized design study is whether two (or more) groups are compared, or a single group is measured over time. These design types differ both conceptually (Valentine & Thompson, 2013) and statistically. Conceptually, QE designs can differ in terms of the logic underlying the counterfactual condition (e.g., a design might estimate an average treatment effect, or a local average treatment effect, see e.g., Imbens & Angrist, 1994). Also a single-group design has no untreated comparison group, implying no placebo effect is expected. The statistical issue that makes multi-group versus single-group effects different is the dependence of measures over time. This causes effect sizes from time-series designs to be distributed differently from effects based on independent groups. This topic has been extensively discussed for effect sizes for continuous outcomes from experimental designs (e.g., Becker, 1988, Gibbons et al. 1993; Morris & Deshon, 2002). More recent work covers multivariate categorical outcomes (Trikalinos & Olkin, 2008, 2012).

A second facet of the nature of the comparison for each study is whether treated and control groups are based on individuals or clusters of individuals. Clusters of individuals in neighborhoods, villages or other units such as medical practices may figure in the provision of treatment, or may simply exist as part of the data collection approach in a primary study. The nesting involved in clustered study designs needs to be considered in the computation and analysis of effect sizes (e.g., Hedges, 2007). The choice of effect size (e.g., what variance is used to standardize a mean difference) will depend on the nature of the desired comparisons.

3.3.2 Group formation. The manner in which groups are formed can impact the distributions and behavior of effect sizes. Recent work on within study comparisons, a method for understanding the conditions under which the results of QEs can approximate the results of RCTs, suggests that researchers must understand, and be able to model, the group formation process (e.g., Cook, Shadish, & Wong, 2008). Thus we concur with Wells et al. that reviewers must record as much detail as possible about how groups are created – by the researcher or via other processes such as self-selection, location, and so forth.

3.3.3 General analytic approach. Because primary studies vary in the ways control is exerted, meta-analysts must document the analytical models used in each QE. The control

variables used – either as design or analysis features – should be examined in the meta-analysis. Lists of theoretically relevant variables or potential confounders/biases in the studies can guide data extraction. DuMouchel (1994) has argued for coding whether particular design features or control variables are missing (rather than present), and whether confounders are present. Coding features as missing allows the intercept in a model of effect size to represent the effect from an ideal study, with other coefficients representing the presence of confounders or biasing features. This approach is in line with Rubin's (1992) concept of predicting out to an ideal study, rather than modeling only the kinds of studies that exist in the literature. Risk of bias tools (e.g., Higgins & Altman, 2008; Higgins et al., 2011; Waddington et al., this issue) and considerations of target studies (e.g., Sterne et al., 2013) may be of use in this task.

Some authors have used risk-of-bias instruments to create scores. Information is scant on the psychometric properties of these instruments (e.g., inter-item correlations, agreement among raters; but see Armijo-Olivo et al., 2012, 2013). The advantage of using such scores as moderators in meta-regression is simplicity (and spare use of data; having a single score leads to a one degree of freedom test). The disadvantage is that different sources of bias in a study may cancel out or compensate for each other; this information is lost when using one single bias score. Further, there is no reason to believe that a single score arising from a quality scale is valid. However, some assessment of study quality is critical, especially because different types of QEs may be differentially at risk for particular biases. For instance, lengthy interrupted time series designs are more susceptible to attrition than one-shot studies. Modeling each potential bias (e.g., with separate predictors for selection bias, attrition, incomplete reporting, etc.) allows for the examination of specific sources of bias. However, it uses more degrees of freedom, requires more studies for modeling in a meta-regression context, and to the extent biases are correlated could suffer from collinearity.

4 WHAT META-ANALYSES OF QES SHOULD INCLUDE

Given the above information, what should the meta-analyst include or report in a synthesis of QEs? As in any meta-analysis the effect sizes (and their variances) must be described. If several types of effects are synthesized together, this choice must be justified. Descriptions (including statistical summaries) of the coded variables and graph(s) displaying the effect sizes are needed.

Authors must examine moderator variables in weighted regressions, analysis-of-variance-like approaches, or Bayesian models. We recommend accommodating for heteroscedasticity via inverse variance weighting (e.g., weighted least squares and generalized least squares) as described below. In the case of within-study dependence (e.g., multiple outcomes per study) generalized least squares analyses (Berkey et al., 1996; Raudenbush, Becker & Kalaian, 1988; among others), cluster-robust standard errors (Hedges, Tipton, & Johnson, 2010), unbalanced panel methods (Rosenberger & Loomis, 2000; Stanley & Doucouliagos, 2012), multi-level models (Kalaian & Raudenbush, 1996) or data reduction may be required.

After this, other typical meta-analytic methods should be applied. We recommend robustness checks and investigations of publication and other biases. Outlier and influence analyses (e.g., eliminating or adjusting certain extreme estimates, or sets of effects) may be especially important. Higgins and Thompson's (2004) permutation test may be useful especially for meta-analyses with small numbers of studies. A discussion of alternative explanations of results is critical because study features and predictor variables tend to be confounded in meta-analyses. Assessment of the generalizability of the conclusions should include a close evaluation of remaining threats/existing weaknesses. Collinearity among the coded features should be assessed (e.g., by cross-tabulating or correlating moderator variables); this may reveal confounding among study features in the literature, and preclude adoption of a single "best explanation" of variation in study effects. Review authors may also find it useful at this point to assess the overall quality of the literature using a system like GRADE (Grade Working Group, 2004; Guyatt et al., 2008) that has been adopted by the Cochrane Collaboration, among others. GRADE views bodies of evidence based on randomized trials to be of the highest quality, but says nothing about how to grade reviews that combine RCTs and QEs. Last, gaps in analyses, or the literature more generally, should be described, along with the inevitable questions for future research.

5 STATISTICAL FRAMEWORK

As noted above a range of possible statistics can serve as effect sizes in QEs. Thus we use a general notation system, with details to be determined by the reviewer's choice of metric. Consider k quasi-experimental studies of a particular topic of interest with comparable effect sizes T_1 through T_k. For simplicity, assume that each study has contributed only one effect size

for one outcome variable. Depending on available study data, the T_i may be regression slopes (b_i), correlations (r_i or some form of partial correlation), standardized mean differences (d_i), or odds ratios. We also assume that all studies contribute the same kind of effect sizes. If quite different types of effects (e.g., *d*'s and *r*'s) have been extracted, they must be expressed on the same metric to be analyzed together.

Assume further that each T_i estimates some population parameter θ_i, and that in large samples the effect size is approximately normal, such that $T_i \sim N(\theta_i, \sigma^2(\theta_i))$. Most meta-analytic effect sizes follow this form (Hedges & Olkin, 1985; Keef & Roberts, 2004). Typically $\sigma^2(\theta_i)$ is a function of the sample size n_i and, for some T_i, of the parameter θ_i. We will denote the sample estimate of $\sigma^2(\theta_i)$ as V_i. Also, we let p represent the full number of features to be considered in a particular analysis, and define x_{ib} as a variable representing study feature b from the i^{th} study.

Given this asymptotic distributional form, many meta-analytic approaches use inverse variance weighting (e.g., via weighted least squares (WLS) or generalized least squares (GLS) analyses; Hedges & Olkin, 1985). Further, in meta-analyses of QEs it is critical to account for the coded design differences and study features discussed above, as well as for any remaining unexplained differences between studies. The model we propose is thus

$$T_i = \beta + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + u_i \ + e_i \,, \qquad (1)$$

where $e_i \sim N(0, V_i)$ represents sampling error in T_i and $u_i \sim N(0, \tau)$ represents variation not explained by the model. Under model 1, the weights $w_i = 1/[V_i + \hat{\tau}]$ are used, and WLS, GLS and Bayes estimation methods may be applied. The x_{ij} reflect study features that may relate to between studies differences in effect sizes. If the u_i term is omitted, model 1 is a fixed-effects model and provides tests of model fit. Omnibus tests of model significance are available under fixed-effects and random-effects models for single or multiple study outcomes. For those who choose to avoid significance testing, all of these methods provide standard errors and confidence intervals. Model 1 is not the only model in contention for use in analyses. Stanley and Doucouliagos (2015, 2016) argue that an unrestricted fixed-effects WLS analysis provides estimates that are statistically comparable to random-effects estimates in all cases and superior to random-effects in many. Other estimation approaches have also been developed, including maximum likelihood as well as Bayesian approaches (e.g., Thompson & Sharp, 1999).

With no predictors, model (1) provides the random-effects estimate of the overall mean. However, it would be rare to stop at a simple random-effects mean, and in some cases that average may be meaningless, especially if effect-size parameters are not consistently defined across studies (e.g., when different factors are controlled across studies).

The *x*'s in the model represent methodological differences (e.g., presence/absence of a given covariate or confound) or other features such as intensity or duration of treatment, and the like. A variety of design characteristics and substantively important treatment or sample features should be analyzed. However, when the number of estimates (studies) is limited and many study features are to be examined, it may not be possible to consider all predictors in one analysis. We suggest prioritizing the most critical predictors and including all of them initially (barring significant issues of multicollinearity). We would give top priority to *x*'s representing biases, controls, and design variations. Differences in the nature of the interventions or populations can capture excess heterogeneity when added as further moderator variables. The final estimated model can be used to obtain the conditional effect size of each design by substituting specific combinations of moderator variable values into the formula in model 1, or used to predict to an optimal result, as suggested by Rubin (1992). In cases of small meta-analyses, key predictors could be run in several sets so as to have fewer predictors per model. Also α levels could be reduced to control for multiple analyses being run, though this would make it harder for slope tests to reach significance.

5.1 Categorical predictors

For categorical moderators, a series of dummy variable *x*'s can be modeled. However, ANOVA-like models (for *p* groups) are often easier to use (e.g., Hedges, 1982). Due to space considerations we do not elaborate on the ANOVA-like model here. A related approach considers categories of studies as panels. Rockers et al. (2015, p. 517, citing Donald & Lang, 2007) identify the fixed-effects panel model as a general QE design for primary research that encompasses both difference-in-difference and control before-and-after QE designs. Stanley and Doucouliagos (2012) have argued that using fixed-effects panel *meta-regression* models gives the systematic reviewer the opportunity to elevate inherently observational meta-analyses to the status of a QE by controlling variations in study quality and other study-level biases or confounders.

5.2 Publication bias

The regression model above has been generalized to accommodate publication bias, by incorporating a predictor with values equal to the standard error or variance of the effect size T_i. Moreno et al. (2009) provide a thorough study of such models including Egger's original method (Egger et al., 1997) in which the standard error or variance is the only predictor. Different weighting schemes have been applied as well. Multiple regressions with substantive or design-based predictors in addition to the standard error or variance appear to investigate potential publication bias *conditional on* other possible sources of heterogeneity (Stanley & Doucouliagos, 2012, 2014). Simulation studies (e.g., Koetse et al., 2010) have shown that meta-regression analysis can simultaneously correct multiple sources of bias in the primary literature, including potential publication selection bias (e.g., Stanley & Doucouliagos, 2016).

6 CONCLUSION

Quasi-experimental studies are likely to contain greater heterogeneity across studies than is typical among RCTs. However, meta-analytic methods are up to this challenge when the relevant research literature contains sufficient information to identify and control statistically the many potential sources of bias and heterogeneity. Finally, the meta-analyst who undertakes the approaches suggested in this article should exercise extreme caution not to make causal statements regarding the relationship between the effect sizes and modeled study features. Relationships between moderators and study outcomes (i.e., the effect size) are observational even when the primary studies in question are randomized (see Thompson & Higgins, 2002).

7 REFERENCES

Aloe AM. An empirical investigation of partial effect sizes for meta-analysis of correlational data. **J Gen Psych** 2014; 141: 47-64.

Aloe AM, Becker BJ, Duvendack M, Valentine JC, Shemilt I, Waddington H (2017). Quasi-experimental study designs for evaluating practice, programs and policies: Collecting data from quasi-experimental studies. **Journal of Clinical Epidemiology** (this issue)

Armijo-Olivo S, Fuentes J, Ospina, M, Saltaji H, Hartling L. Inconsistency in the items included in tools used in general health research and physical therapy to evaluate the methodological quality of randomized controlled trials: A descriptive analysis, **BMC Med Res Meth** 2013; 13: 116.

Armijo-Olivo S, Stiles CR., Hagen NA, Biondo PD, Cummings, GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. **J Eval Clin Prac** 2012; 18: 12–18. doi: 10.1111/j.1365-2753.2010.01516.x

Bärnighausen T, Oldenburg C, Tugwell P, Bommer C, Ebert C, Barreto M, Djimeu E, Haber N, Waddington H, Rockers P, Sianesi B, Bor J, Fink G, Valentine J, Tanner J, Stanley T, Sierra E, Tchetgen Tchetgen E, Atun R, Vollmer S (2017). "Quasi-experimental study designs for evaluating practice, programs and policies: assessing the assumptions." **Journal of Clinical Epidemiology** (this issue).

Becker, B. J. Influence again: An examination of reviews and studies of gender differences in social influence. In: Hyde JS, Linn MC, eds, **The psychology of gender: Advances through meta-analysis**, 1986; Baltimore, MD: Johns Hopkins Press.

Becker, BJ. Synthesizing standardized mean-change measures. **Brit J Math Stat Psych** 1988; 41: 257-278.

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 BJ. Multivariate meta-analysis. In: Tinsley HEA, Brown SD, eds; **Handbook of applied multivariate statistics and mathematical modeling,** 2000; San Diego, CA: Academic Press.

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.

Becker BJ, Aloe AM, Olkin I. Dependence of slopes from a single sample. **J Educ Behav Stat** under review.

Berkey CS, Anderson JJ, Hoaglin DC. Multiple-outcome meta-analysis of clinical trials. **Stat Med** 1996; 15(5), 537-557.

Berlin JA. Invited commentary: Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. **Am J Epi** 1995; 142(4), 383-387.

Bryant FB, Wortman PM. Methodological issues in the meta-analysis of quasi-experiments. **New Dir Prog Eval** 1984; 24: 5–24.

Campbell DT, Stanley JC. **Experimental and Quasi-Experimental Designs for Research.** Chicago: Rand McNally, 1963.

Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. **J Policy Anal Mgmt** 2008; 27: 724-750.

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

Donald SG, Lang K. Inference with difference-in-differences and other panel data. **Review of Economics and Statistics** 2007; 89(2): 221-233.

DuMouchel, W. **Hierarchical Bayes Linear Models for Meta-analysis** (Technical Report No. 27). Research Triangle Park, NC: National Institute of Statistical Sciences, 1994.

Dunning, T. **Natural experiments in the social sciences: A design-based approach**. Cambridge, UK: Cambridge University Press, 2012.

Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. **BMJ** 1997; 315: 629-634.

Gibbons RD, Hedeker DR, Davis JM. Estimation of effect size from a series of experiments involving paired comparisons. **J Ed Stat** 1993; 18(3), 271-279.

Glanville J, Eyers J, Jones AM, Shemilt I, Wang G, Johansen M, Fiander M, Rothstein H (2017). Quasi-experimental study designs for evaluating practice, programs and policies: Identifying quasi-experimental studies to inform systematic reviews. **Journal of Clinical Epidemiology** (this issue).

GRADE Working Group. Grading quality of evidence and strength of recommendations. **BMJ** 2004; 328: 1490-1494.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. **BMJ** 2008; 36: 924-926.

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

Hedges LV, Olkin, I. **Statistical Methods for Meta-analysis**. New York, NY: Academic Press, 1985.

Hedges LV., Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. **Res Syn Meth** 2010; 1(1):39-65.

Higgins JPT, Altman DG. Assessing risk of bias in included studies. In: Higgins JPT, Green S, eds. **Cochrane Handbook for Systematic Reviews of Interventions.** Wiley, 2008:187-241.

Higgins JPT, Altman DG, Gotsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF Weeks L, Sterne JAC. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. **BMJ** 2011; 343: d5925. doi: 10.1136/bmj.d5928

Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. **Stat Med** 2004; 23(11): 1663-1682.

Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. **Econometrica** 1994; 62(2): 467-475.

Kalaian, HA, Raudenbush SW. A multivariate mixed linear model for meta-analysis. **Psych Meth** 1996; 1(3): 227-235.

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

Koetse MJ, Florax, RJGM, de Groot HLF. Consequences of effect size heterogeneity on meta-analysis: A Monte Carlo experiment, **Stat Meth Appl** 2010; 19, 217–36.

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 JN, Bärnighausen T, El-Jardali F (2017). 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).

Lipsey MW. Those confounded moderators in meta-analysis: Good, bad, and ugly. **Annals Am Acad Pol Soc Sci** 2003; 587 (1): 69-81.

Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, Cooper NJ. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. **BMC Med Res Meth** 2009; 9:2 doi:10.1186/1471-2288-9-2.

Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. **Psych Meth** 2002; 7(2): 105-125.

Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. **Psych Bull** 1988, 103(1): 111-120.

Reeves BC, Wells GA, Waddington H (2017). Quasi-experimental study designs for evaluating practice, programs and policies: Classifying studies evaluating effects of health interventions." **Journal of Clinical Epidemiology** (this issue).

Richardson WS, Wilson MC, Nishikawa J, Hayward, Robert SA. The well-built clinical question: A key to evidence-based decisions. **ACP Journal Club** 1995; 123, A-12.

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

Rockers PC, Tugwell P, Grimshaw J, Oliver S, Atun R, Røttingen J, Fretheim A, Ranson MK, Daniels K, Luiza VL, Bärnighausen T (2017). 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).

Rosenberger RS, Loomis JB. Panel stratification in meta-analysis of economic studies: An investigation of its effects in the recreation valuation literature, **J Agr App Econ** 2000; 32: 459-470.

Rubin DB. Meta-analysis: Literature synthesis or effect-size surface estimation? **J Educ Stat** 1992; 17(4): 363-374.

Slavin, RE. What works? Issues in synthesizing educational program evaluations. **Ed Researcher** 2008; 37: 5-14.

Stanley TD, Doucouliagos H. **Meta-Regression Analysis in Economics and Business,** Oxford: Routledge, 2012.

Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. **Res Syn Meth**. 2014; 5:60-78.

Stanley TD, Doucouliagos H. Neither fixed nor random: Weighted least squares meta-analysis. **Stat Med** 2015; 34: 2116–27.

Stanley TD, Doucouliagos H. Neither fixed nor random: Weighted least squares meta-regression. **Res Syn Meth** 2016; DOI: 10.1002/jrsm.1211.

Sterne J, Higgins JPT, Reeves B. **Extending the Cochrane Risk of Bias tool to assess risk of bias in *non-randomized studies.*** Draft for circulation provided by Peter Tugwell, 2013.

Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? **Stat Med** 2002; 21(11):1559-73.

Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: A comparison of methods. **Stat Med** 1999; 18: 2693-2708.

Trikalinos TA, Olkin I. A method for the meta-analysis of mutually exclusive binary outcomes. **Stat Med** 2008; 27(21): 4279–4300.

Trikalinos TA, Olkin I. Meta-analysis of effect sizes reported at multiple time points: A multivariate approach. **Clin Trials** 2012; 9(5): 610-620.

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.

Van Den Noortgate W, López-López J, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. **Beh Res Meth** 2013; 45: 576-594.

Waddington H, Aloe A, Becker BJ, Djimeu, E, Reeves B, Tugwell P. Risk of bias assessment in credible quasi-experimental studies. This issue.

Wells GA, Shea B, Higgins JPT, Sterne J, Tugwell P, Reeves BC. Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. **Res Syn Meth** 2013; 4(1): 63-77.

Wortman PM. Lessons from the meta-analysis of quasi-experiments. In: Bryant FB, Edwards J, Tindale RS, Posavac, EJ, Heath L, Henderson-King E, Suarez-Balcazar Y, eds. **Methodological Issues in Applied Social Psychology**, 1992; Social Psychological Applications to Social Issues, Volume 2, 65-81.

ACKNOWLEDGEMENTS

Betsy Becker's contribution to this research was supported by a grant from the National Science Foundation (NSF DRL-1252338). Ariel Aloe's contribution to this research was supported by a grant from the National Science Foundation (NSF DRL-1252263).

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.