Using Systematic Reviews and Meta-Analyses to Inform Public Policy Decisions

Jeffrey C. Valentine
University of Louisville

Spyros Konstantopoulos
Michigan State University

Commissioned article prepared for the Committee on the Use of Economic Evidence to Inform Investments in Children, Youth, and Families

The National Academies of Sciences, Engineering and Medicine

Article prepared for the Committee on the Use of Economic Evidence to Inform Investments in Children, Youth, and Families. For more information see:
http://sites.nationalacademies.org/DBASSE/BCYF/EconForKids/index.htm

This paper is intended as a discourse on the characteristics that systematic reviews and meta-analyses should have in order for them to provide a useful basis for informing public policy (including, but not limited to, benefit-cost analyses and program adoption decisions). Before we begin, a note about terminology is warranted. In common social science conversation, the term “meta-analysis” is used broadly. Many scholars (e.g., Higgins & Green, 2011; Cooper, Hedges, & Valentine, 2009) prefer to distinguish between a “systematic review” and a “meta-analysis”. The term systematic review is used to describe state-of-the-art literature searching, data extraction, and study quality assessment techniques. We use the term meta-analysis to refer to the quantitative synthesis of the results of multiple studies. Meta-analyses are not necessarily based on systematic reviews (though in most cases they should be!), and not all systematic reviews culminate in a meta-analysis.

After a brief overview of the processes involved in a systematic review, we address three broad points. First, systematic reviews ought to be based on a thorough literature search for potentially relevant studies. Second, studies considered for inclusion in a systematic review should be subjected to thorough quality appraisal. Finally, integrating evidence from randomized and nonrandomized experiments requires careful consideration of the required assumptions and tradeoffs involved.

SYSTEMATIC REVIEWING

As a research enterprise, systematic reviews are structured like any other empirical study. That is, a research question has to be formulated, and data have to be collected, analyzed, interpreted, and reported. In fact, one useful way of framing a systematic review is to say that it is a form of survey research, but instead of surveying people, research studies are surveyed (Lipsey & Wilson, 2001).

Systematic reviews represent an attempt to minimize bias in the collection, evaluation, and synthesis of studies on a particular research question. Though relatively rare in everyday academic practice, most experts recommend, and some organizations implement, the use of protocols as a primary means of achieving the bias reduction aim (e.g., the Campbell Collaboration, the Cochrane Collaboration, and the U.S. Department of Education’s What Works Clearinghouse). Protocols are documents that specify, in as much detail as possible and in advance of actually doing the work, the methods that will be employed in the systematic review and meta-analysis[1]. It has become increasingly clear that study authors systematically make analytic decisions that are biased against the null hypothesis (e.g., Gøtzsche, 2006; Simonsohn, Nelson, & Simmons, 2014). A highly operational protocol helps protect the consumers of systematic reviews, and the systematic reviewers themselves, from data-driven decisions that serve to advance the interests of the systematic reviewers. An additional benefit of a protocol is that is forces synthesists to think deeply, before data collection, about important aspects of the review. For example, it is likely that interventions will be implemented at least slightly differently across studies. Some might last longer (say 6 months) while others might be shorter (say, one month). Or some might be aimed at preschoolers, while others are aimed at early

[1] By way of example, the Cochrane Collaboration’s protocol guidelines can be found here: http://ph.cochrane.org/sites/ph.cochrane.org/files/uploads/Guide%20for%20PH%20protocol_Nov%202011_final%20for%20website.pdf.

elementary school aged students. Decisions need to be made regarding boundary conditions – i.e., what characteristics must an intervention have (or alternatively, cannot have) in order to be included in the review. Usually there are no simple answers to these questions, and in fact in our experience, these boundary questions are among the more interesting aspects of doing a systematic review. It should be clear that deep substantive expertise is needed in order for the review team to reach defensible decisions. Consider the case of studies that are aimed at students of different ages. The researchers need to be able to make a judgment about whether the intervention's effects are likely to differ across these populations, or if they are expected to be similar enough that they can be combined. We note here – and reinforce later – that the number of studies that are ultimately expected to be included in the review is an important additional consideration. If the expected number of studies is large, then it makes more sense to be inclusive if one is unsure about the boundary conditions. Inclusivity is defensible in this case because the large number of studies will allow the reviewers to conduct moderator, subgroup, or sensitivity analyses, and perhaps to control for study characteristics in the analysis. If the expected number of studies is small, then there will be few or no such analysis opportunities, making inclusion decisions much more consequential.

Searching the Literature

The backbone of any systematic review is the literature search. It should be clear that, generally speaking, if we want to know about the effects of an intervention on an outcome, we are better off having access to all of the studies that have been conducted on that intervention rather than to only a selected portion of that literature. Including different sets of studies in a synthesis can lead to different conclusions about the effects of a program or practice. Furthermore, the most defensible conclusions can only be drawn from having the most complete set of available research that it is possible to assemble. One major challenge to identifying this comprehensive set of studies arises from publication bias, which refers to the tendency of the published literature to suggest effects that are larger than those observed in unpublished studies. Publication bias occurs because in many disciplines study authors are less likely to submit for publication, and journal editors and peer reviewers are less likely to accept, studies that do not have statistically significant findings on their primary outcomes. These behaviors are attributable in part to a persistent disregard for statistical power, and to common misconceptions regarding the interpretation of probability values arising from null hypothesis significance tests. Under the assumption that standard errors in small sample studies are similar in magnitude, statistically significant estimates are probably larger than non-statistically significant estimates. Because larger estimates are more likely to be published, this results in an evidence base that suggests effects that are larger than they really are, and therefore in a bias against the null hypothesis.

In light of these concerns, to be a useful decision making aid a systematic review should involve a robust literature search, the aim of which is to identify all relevant studies that have been conducted. The literature search will ideally be developed by researchers with expertise in the substantive aspects of the research question (e.g., the different ways that the intervention might be described across academic disciplines and policy contexts) and in literature retrieval; specially trained librarians are very helpful in this regard. Furthermore, multiple channels should be searched to supplement the leads obtained from academic databases. For example, many authorities recommend using a "snowball" technique, which involves writing the authors of relevant studies to see if they have other studies that the search has not uncovered (e.g., studies that were never submitted to a journal). Relevant advocacy and policy organizations are also potentially good sources for material, as are government and nonprofit agencies that fund related research.

There are two natural consequences of a limited search: a more limited evidence base and a meta-analytic effect size estimate that is probably biased against the null hypothesis. The evidence base is likely to be limited in both obvious and subtle ways. For example, a limited literature search will likely uncover fewer studies for review than a comprehensive one, and less obviously will likely over-represent the authors' professional networks (e.g., will be less likely to uncover eligible studies conducted in different academic disciplines). In other words, the literature search should include published and unpublished work in various academic disciplines to enhance the external validity of the synthesis' results. As such systematic reviews not based on a vigorous literature search should be viewed skeptically.

Study Quality Appraisal

Empirical studies vary in terms of the rigor with which they are conducted, and the quality of the studies comprising any research synthesis can have an impact on the validity of conclusions arising from that synthesis (e.g., estimates from well-designed and executed experiments have higher internal validity). The wide agreement on these points is reflected in the fact that virtually all synthesis efforts invoke validity when making decisions about (a) what to include in their syntheses and (b) how to draw distinctions between the studies that meet criteria for inclusion. That agreement aside, the actual mechanics of conducting quality assessments vary dramatically. In part, the variation in quality assessment procedures is attributable to the relative paucity of empirical evidence that can be used to inform the development of a quality assessment scheme. The relative lack of evidence is made worse by the near certainty that much of what constitutes "study quality" is dependent on the context of the research question: A threat to validity in one context might not be a threat to validity in another context. This means that even when empirical evidence is available, its applicability to different contexts is highly uncertain. For example, most quality assessment schemes involve an assessment of attrition (i.e., the loss of participants from the time of assignment to the time of outcome measurement). However, Valentine and McHugh (2007) demonstrated that in studies of routine educational practices in elementary school attrition is likely exogenous to the intervention: Whether a given student is present or absent for a posttest appears to be much more likely to be due to chance (i.e., the child just happened to be ill that day) than it is to be related to anything related to the intervention. That is, in this case differential attrition in treatment and control groups is likely to be random and as such would not bias the results. Therefore, the quality of a study of routine educational practices in elementary school might be improperly ranked or categorized on a quality assessment scheme that includes attrition, possibly leading to an inappropriate down weighting of the evidence from that study.

Guidance on how to develop a quality assessment scheme is inconsistent, as is the guidance on the specific elements of study method and procedure to include in such an assessment system. That said, there are certain quality assessment practices that we can safely say ought to be avoided. Chief among these is using of scales that sum to a single score representing quality (Valentine & Cooper, 2008). As an example of the problems evident in these scales, Jüni, Witschi, Bloch, and Egger (1999) identified 25 quality scales, 24 of which had been published in peer-reviewed medical journals, and applied them to a set of studies investigating the effects of a new drug aimed at preventing post-operative blood clots. Jüni et al. (1999) found that the results of the meta-analysis depended on which quality scale was used to evaluate the studies. For about half of the quality scales, there was no difference in effects between the "high quality" and the "low quality" studies. However, about one quarter of the quality scales yielded results suggesting that the new treatment was superior to standard treatment in the high quality studies, but not in the low quality studies. For the remaining quarter of quality scales, this pattern was reversed. That is, in the high quality studies the effects of the new and standard treatment were not different, but in the low quality studies the new treatment was superior to the standard treatment.

The dependence between the outcomes of the synthesis and the quality scale used is problematic and indicates lack of robustness (or model dependency) in the results. Imagine that a sophisticated team of physicians choose one quality scale, use it to exclude the low quality studies, and the base treatment decisions on what the high quality studies suggest about the effectiveness of the drug. The fact that treatment choice is at least somewhat dependent on the specific quality scale chosen is disconcerting, as is the fact that another team of physicians' choice of a different study quality scale could lead them to different treatment decisions. By and large, the problem illustrated here is endemic to virtually all quality scales, and therefore the use of quality scales – especially those that result in a single score – should be avoided (see Valentine & Cooper, 2008, for a thorough review of the problems shared by most quality scales).

How should variation in study quality be addressed in a systematic review? Some have argued for the inclusion of study quality weights in meta-analysis. However, our review of the work by Jüni and colleagues should make the difficulty of doing this clear: one needs valid weights, and none exist. In fact, Ahn and Becker (2011) both demonstrated how one could incorporate study quality weights into a meta-analysis and cautioned against doing so, writing "because quality weights lead to bias in almost every condition studied, we recommend against the use of quality weights" (p. 555). Instead we believe that best practice - and therefore the best basis for building policy - is first to identify a very small number of highly defensible design features that studies are required to have. In a systematic review of interventions, for example, the synthesists might have good reason to believe that assignment to conditions should be done randomly, because a body of research suggests that nonrandom assignment results in biased estimates (we will have more to say on this point later). The job training literature is one research area in which this assertion may be true (e.g., Fraker & Maynard, 1987; Glazerman et al., 2003). The approach of identifying a very small number of highly defensible design features that studies are required to have is consistent with the approach taken by the Cochrane Collaboration. In most of their systematic reviews, one inclusion criterion is that the study used randomization to place participants into groups. Then, after identifying a very small number of highly defensible design features (e.g., random assignment) that studies are required to have to be eligible, authors should identify other dimensions of study quality likely to be relevant for their research question. Again, this recommendation is consistent with the Cochrane Collaboration's approach. Research teams reviewing for Cochrane rate studies according to the "Risk of Bias" tool, which addresses eight additional study quality dimensions (e.g., whether the randomization scheme was successfully carried out; whether attrition was a threat to the validity of the study's conclusions).

Depending on whether the number of studies in the meta-analysis is large or not, authors have different recommended courses of action. First, if the number of studies in the meta-analysis is large, authors can investigate the extent to which the quality indicators covary with effect sizes using meta-regression (a form of weighted regression), and can generate impact estimates that control for these dimensions. If the number of studies is not large enough to support meta-regression, univariate tests of study quality dimensions may be feasible, though power can be low for these tests (see Hedges & Pigott, 2004; Valentine, Pigott, & Rothstein, 2010). Alternatively, separate meta-analyses could be carried out, grouping studies by some critical dimension (e.g., RCTs in one group, nonequivalent control group designs in another). If the number of studies is too low to allow for any statistical investigation of the extent to which study quality indicators covary with effect size, authors should at least provide a rich description of each included study on the selected dimensions.

All of this raises the question: What other dimensions should be examined? In public policy contexts, most discussions about study quality center on internal validity of results, i.e., the believability of the causal claim; virtually all quality scales assess this dimension. However other considerations are also important. For example, measurement issues receive less attention than they should and indeed, measurement in the social sciences is often an embarrassment (as exemplified by the ease with which one can find examples of study authors asserting that a measure is reliable and valid by providing a citation to an external source without much consideration for whether the measure is valid for the purpose for which the authors are using it, and whether the scores produced by the measure - given the sample and the context - are reasonably reliable). Additional concerns have been raised regarding other dimensions. For example, studies with high levels of intervention developer involvement have been demonstrated to obtain larger effect sizes, on average, than studies with little or no developer involvement (e.g., Lipsey, Landenberger, & Wilson, 2007; Petrosino & Soydan, 2005). While it should be clear that it is best to distance individuals with a strong stake in an outcome from data analysis decisions, it is also important to note that these concerns should not be taken as an indication that developers are not to be trusted, or that studies with high developer involvement ought to be automatically discounted. For example, the larger effect sizes obtained with high developer involvement could be a function of the fact that program developers are better implementers of their own interventions (Valentine, et al., 2011). Still, an indicator about whether developers were involved in data analyses or not would allow synthesists to control for this potential confounding either in a meta-regression (e.g., use developer involvement as a predictor/moderator) or in subgroup analyses (e.g., combine estimates with no developer involvement separately).

In summary, the methods reviewers should use to address variations in study quality depend on the number of studies the reviewers expect to be able to include in their analyses. With a large number of studies meta-regression can be used to investigate quality indicator and effect size covariation, and to control for quality indicators in effect size estimation. Unfortunately most reviewers find themselves in information poor environments (for example, the median number of studies in What Works Clearinghouse intervention reports is about 2.5). In cases like this the choices are not nearly as good. It is clear that most study quality scales (especially those that arrive at a single number that represents study quality) and the use of study quality weights should be avoided. Therefore, when the number of studies is expected to be small, prior to data collection reviewers should make and carefully justify decisions about study characteristics that will result in study inclusion and exclusion, and then carefully describe the status of the included studies on other important study dimensions (such as other study quality indicators).

Synthesizing the Evidence

Once a set of studies that are to be included in a review have been identified, the next step is to synthesize the studies. In the context of a literature review a synthesis can be broadly defined as a conclusion or summary about what the literature reveals about the research question. While there are many methods for synthesizing the results of studies, we focus on three forms of meta-analysis: classical fixed effect meta-analysis, classical random effects meta-analysis, and Bayesian approaches to meta-analysis. The approaches have somewhat different sets of assumptions, strengths, and limitations. Our discussion assumes that the research designs to be synthesized are founded on similar inferential logic. RCTs and single group interrupted time series designs, for example, rely on fundamentally different ways of arriving at an inference about the effectiveness of an intervention, and therefore cannot be combined in a way that results in a sensible meta-analytic mean.

Classical fixed effect meta-analysis

Classical meta-analyses usually use an inverse variance weighting scheme; doing so gives proportionally more weight to larger-sample studies. Classical fixed effect meta-analysis starts with the assumption that all of the studies are estimating the same population parameter. One way to think about this assumption is that if all of the studies included in the review were infinitely large, they would all have the same effect size. In other words, studies yield different effect sizes only because of sampling error, and the weights are simply a function of sample size. Fixed effect assumptions are most applicable when the studies included in a review are very close replicates of one another.

Classical random effects meta-analysis

Rather than assuming that all studies are estimating the same population parameter, classical random effects meta-analysis starts with the assumption that the sample estimates are sampled randomly from a population distribution of effects that are centered on some mean. Thus, study effects are expected to differ from one another both due to sampling error and due to observed and unobserved differences in study characteristics. The weights used reflect this difference: rather than just being a function of sample size, the weights are a function of sample size and an estimate of the between-studies variation. As a result, the confidence intervals arising from a random effects meta-analysis can never be smaller, and will often be larger than their fixed effect counterparts (unless the between-studies variation is exactly zero). Due to the commonly-observed negative correlation between sample size and effect size effect size estimates are often somewhat larger in random effects models (this is due to the fact that random effects weights are more equal than fixed effect weights – in cases of extreme heterogeneity, the random effects may be essentially the same for all studies regardless of sample size). The increased effect size is usually not large enough to offset the increase in the standard error. As such statistical power is usually lower in random effects models, relative to that observed in fixed effect models.

Though occasionally meta-analysts present the results of both fixed effect and random effects models, generally one is chosen (preferably on grounds established *a priori*). A limitation of the fixed effect model is the assumption that all studies are estimating the same population parameter. Given that in most cases the studies included in a review will vary from one another in multiple known and unknown ways, this seems like a difficult assumption to entertain. Therefore the random effects model likely provides a better fit to most meta-analyses. However, as discussed random effects meta-analysis requires an estimate of the between studies variance, and this estimate is often poor if the number of studies is not large (at a bare minimum, at least five studies are needed for a reasonably precise estimate, though in most cases more studies – at least 10 - will be needed). A poor between-studies variance estimate can result in estimated effects that are either too small or too large, and standard errors for those effects that are also either too small or too large (i.e., inferential tests can be either biased against the null hypothesis or biased in favor of it), and there is no way to either predict the direction of the bias or to assess the extent of it.

Bayesian meta-analysis

A Bayesian meta-analysis addresses the problems associated with classical fixed effect and random effects meta-analysis by bringing prior knowledge to bear on the research question. In fact, one way to think about Bayesian statistics is as a collection of techniques that synthesize what is believed to be known prior to the start of a study (known as a prior) and the study results. The priors chosen can be weak or strong (sometimes couched in terms of “uninformative” vs. “informative” priors), and the data can be weak or strong (the intervention’s effect can be estimated imprecisely or precisely). A Bayesian synthesis is weighted in the sense that if the data are strong, then even an informative prior will tend not to have much influence on the analysis. However, if the data are weak, then the prior can have a larger influence on the results.

Priors can be based on anything, including subjective opinion, and many classical statisticians worry that the results of Bayesian analyses are dependent on the essentially arbitrary priors chosen. To be a good basis for public policy the priors chosen should be defensible. There are many ways to arrive at defensible priors. First, Press (2003) argues that the most defensible priors are those that reflect a position of ignorance. Such priors are often referred to as “uninformative”. With uninformative priors, the data will tend to dominate, and indeed, the results of a Bayesian random effects meta-analysis based on uninformative priors will generally be fairly similar to the results of a classical random effects meta-analysis. Another option is to elicit expert opinion (see Thompson et al., 2011, for an example of how this might be done). Finally, priors can be informed by empirical evidence. For example, since 2010 the Institute of Education Sciences’ What Works Clearinghouse has conducted over 60 meta-analyses of interventions in the preK-12 educational context. The results of these provide a reasonable basis for priors. Within the Bayesian framework sensitivity analyses are also informative and should be conducted by synthesists. That is, analysts could use various uninformative or informative priors that seem reasonable to test the robustness of results.

In summary, from a classical statistics standpoint the random effects model is conceptually the best fit. But an important statistic (the between-studies variance) is poorly estimated if there are few studies, as is common in public policy contexts. As such the classical fixed effect model might be an alternative. However, the classical fixed effect model is often not a good conceptual fit because it invokes the assumption that the studies are all assessing the same population parameter (i.e., that they are close replicates of one another). In addition, in a fixed effects model the generality of findings is restricted to the sample of studies at hand, which to some degree undermines the advantage of a meta-analysis. A Bayesian approach addresses the main challenge of employing the random effects model by bringing in additional information in the form of the prior.

Synthesizing randomized and nonrandomized evidence

For syntheses of intervention research, an important consideration in many public policy contexts is (a) whether nonrandomized evidence should be included in the review and if so (b) whether nonrandomized evidence should be meta-analyzed with evidence arising from randomized studies. There is good reason to prefer evidence from randomized experiments; chief among these is that its assumptions are transparent. That said, nonrandomized studies are common, and in many policy contexts evidence is hard to come by. Furthermore, in many disciplines experiments tend to be quite different – and not in good ways – from nonrandomized studies. For example they tend to be conducted on more selected populations, to have implementation more heavily monitored, and to have more involvement of the developer (see Lipsey, et al., 2007).

Due to these considerations, scholars have devoted considerable energy to assessing the conditions under which nonrandomized studies might approximate the results of randomized experiments. A line of this research is known as “within study comparisons”. For example, Shadish et al. (2008) randomly assigned college students to be in either an RCT or in a nonrandomized experiment. Students in the RCT arm of the study were randomly assigned to be in either a vocabulary or in a math intervention. Students in the non-randomized arm were allowed to self-select into the vocabulary or the math intervention. The effects of training were assessed by giving participants a math and a vocabulary posttest. For the vocabulary treatment effect, the scores of students in the vocabulary condition on the vocabulary posttest were compared to the scores of students in the math condition on the vocabulary posttest (and vice versa for the math treatment effect). Overall, ANCOVA based on a rich covariate set reduced bias about 94% for the vocabulary training effect and 84% for the math training effect. This finding suggests that if initial bias is 0.20 standard deviations, the bias remaining after controlling for a rich set of well-measured covariates is only about 0.02 standard deviations.

In addition to Shadish et al. (2008), a number of other studies using this method have been conducted in educational contexts (Wong et al., 2015, identified 13 such studies in education). Though this research is still emerging and is complex, in general, these studies suggest three findings. First, controlling for well-measured baseline covariates has the potential to substantially reduce bias. A pretest of the outcome is likely the single most effective covariate to consider. For outcomes without a good, well-measured pretest, attempting to model the selection process that unfolded in the studies or including covariates from a variety of domains seem to be the most promising approaches. Thus, careful thinking about selection processes and measuring important covariates to reduce or eliminate confounding constitute crucial components in designing empirical studies, and most nonexperimental studies probably do not reach this goal. Second, there may be good reason to prefer “local” comparison groups to comparison groups drawn from non-local contexts, especially for situations in which (a) no good pretests exist and (b) the selection model operating in the studies is unknown or differs across studies (leading to the use of covariates measured across a variety of domains). Local comparison groups have a higher likelihood of being on average similar to local treatment groups which facilitates causal arguments. This is one possible explanation for the previously-mentioned preference for randomized experiments in the job training literature, as studies in that domain are less likely to use local treatment groups. Finally, relying on commonly collected and easily measured covariates (such as basic demographics) is unlikely to substantially reduce bias. For that reason estimates produced from empirical analyses of preexisting secondary datasets (e.g., NAEP, TIMSS, etc.) may not be unbiased because of lack of important observed covariates that can affect the association between independent and dependent variables.

If the first challenge associated with including randomized and nonrandomized experiments in the same meta-analysis is associated with the believability of the nonrandomized estimate, a second is associated with the comparability of the estimates arising from randomized and nonrandomized studies. We mentioned that one benefit of an RCT is the transparent nature of its assumptions. That is, because of equal chance of being in a treatment or a control group, on average, treatment and control groups are equivalent (within the limits of sampling error) on all observed and unobserved variables at baseline. This transparency translates into appropriately simple statistical tests (such as an independent samples *t*-test for a two group experiment). Although not a requirement, analyzing data from experiments using multiple regression analyses that include covariates with predictive power is common practice in the social sciences because it results in more precise estimation of effects (i.e., smaller standard errors). The analysis of nonrandomized experiments will almost always require a somewhat more sophisticated analytic approach, usually using some form of multiple regression analysis or the related analysis of covariance. There are two resulting difficulties with estimates of regression models however. First, the specific models generating the estimates will usually differ across studies. That is, it is likely that no two studies will use exactly the same set of covariates. This means that the interpretation of the estimates is at least a little different in every study. The second difficulty is related to the first. Since the models differ across studies, the standard errors of these estimates will also differ. Eventually, this translates to different kinds of weights across studies because an inverse function of the standard errors is typically used to construct weights. There is no easy way to address these difficulties. One way to synthesize quantitatively estimates from regression models across studies is to first ensure that the estimates are on the same scale and then to combine estimates from regression models that seem comparable (i.e., use very similar if not same covariates). That is, one could combine estimates from regression models that have included the same important covariates in their specifications. In education, such covariates are prior achievement scores, SES, and ethnicity.

Still, it is unclear that combining experimental and non-experimental estimates is the right practice, unless at least two important conditions have been met: (a) non-experimental studies have measured important covariates (e.g., prior measures) that eliminate or reduce dramatically confounding, and (b) quasi-experimental methods that facilitate causal inference have been employed. For example, propensity score methods can produce valid causal estimates when a rich set of well-measured covariates is available (Rosenbaum, 2002). In addition, regression discontinuity designs approximate a localized experiment around the discontinuity points and thus can produce causal estimates that are comparable to those obtained from experiments (Shadish, Cook, & Campbell, 2002). Estimates produced from studies that have utilized quasi-experimental designs or methods that facilitate causal inference and have controlled for important well-measured covariates should be very similar to experimental estimates. In that case combining experimental and non-experimental estimates may be warranted. Nevertheless, when meta-analysts decide to combine estimates from experiments and non-experiments they need to be transparent about their rationale and they need to defend their decisions, methods, and analytic modeling.

CONCLUSION

In order to achieve a minimum threshold of validity of estimates synthesists and meta-analysts would benefit from the following guidelines. First, we recommend a thorough/exhaustive literature search on the topic of interest. This indicates that all possible studies (published or not) in that area should have an “equal” chance of being in the sample of studies. This will guarantee to a high degree the external validity of the evidence produced by syntheses and meta-analyses. Second, reviewers should establish and defend clear inclusion criteria prior to data collection (e.g., via a review protocol). Third, a well-thought out, rigorous, robust, and appropriate quality appraisal of the studies in the sample is necessary. The quality appraisal can be used to investigate the relationship between various study quality indicators and effect size, and to control for study quality characteristics if the number of studies is large enough to support meta-regression, or to describe study variation along important dimensions if not. Fourth, synthesists should construct appropriate indexes that ensure comparability of estimates across studies. That is, estimates from various studies should be on the same scale and thus mean the same thing. Fifth, methods and statistical analyses need to be carefully chosen. Different analyses could be explored to test the robustness of the estimates. Finally, combining regression estimates across studies that have used different designs (e.g., experiments or non-experiments) and different specifications (different set of covariates) requires thorough thinking. The selected methodology and analytic modeling should be transparent and well reasoned. Syntheses of experimental and non-experimental estimates should be clearly justified and coherent and explicit arguments should be made about the comparability of estimates across studies.

REFERENCES

Ahn, S., & Becker, B. J. (2011). Incorporating quality scores in meta-analysis. *Journal of Educational and Behavioral Statistics, 36*(5), 555-585.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources, 22*(2), 194-227.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy, 589*, 63-93.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological methods, 9*(4), 426.

Higgins, J. P. T. (2011). Green S. Cochrane handbook for systematic reviews of interventions version 5.1. 0. *The Cochrane Collaboration, 5*.

Lipsey, M., Landenberger, N. A., & Wilson, S. J. (2007). Effects of cognitive-behavioral programs for criminal offenders: A systematic review. *Campbell Systematic Reviews*, *3*(6).

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology*, *1*(4), 435-450.

Press, S. J. (2009). *Subjective and objective Bayesian statistics: principles, models, and applications* (Vol. 590). John Wiley & Sons.

Rosenbaum, P (2002). *Observational Studies (2nd edition)*. New York: Springer.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi- experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Thompson, S.G., Ekelund, U., Jebb, S.A., Lindroos, A.K., Mander, A.P., Sharp, S.J., Turner, R.M., Wilks, D.C. (2011). A proposed method of bias adjustment methods for meta-analyses of published observational studies. *International Journal of Epidemiology, 40*:765–777.

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., Kellam, S., Mościcki, E. K., & Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12*, 103-117.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, *13*, 130-149

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215-247.