Journal of Methods and Measurement in the Social Sciences,
Vol.4, No.2, 21-45, 2013

Complementary Meta-Analytic Methods for the Quantitative Review of Research: 1. A Theoretical Overview

Aurelio José Figueredo, Candace Jasmine Black, and Anne Grete Scott
University of Arizona

Contents Meta-Analysis is a procedure designed to quantitatively analyze the methodological characteristics in studies sampled in conventional meta-analyses to assess the relationship between methodologies and outcomes. This article presents the rationale and procedures for conducting a *Contents* Meta-Analysis in conjunction with conventional *Effects* Meta-analysis. We provide an overview of the pertinent limitations of conventional meta-analysis from methodological and meta-scientific standpoint. We then introduce novel terminology distinguishing different kinds of complementary meta-analyses that address many of the problems previously identified for conventional meta-analyses. We would also like to direct readers to the second paper in this series (Figueredo, Black, & Scott, this issue), which demonstrates the utility of Contents Meta-Analysis with an empirical example and present findings regarding the generalizability of the effect sizes estimated.

Keywords: meta-analysis, methods, scientific progress, contents meta-analysis, metascience

We propose a new meta-analytic procedure, which we call *Contents* Meta-Analysis, for the purpose of analyzing the methodological characteristics of studies prior to conducting a conventional meta-analysis. The objective of this article is to present the rationale and detail the procedures for conducting a Contents Meta-Analysis in conjunction with the conventional procedures that we call *Effects* Meta-analysis, to distinguish it more clearly from the former. Our argument is structured to first provide an overview of the limitations of conventional meta-analysis from methodological and meta-scientific standpoint, and then to introduce some novel terminology distinguishing these different kinds of meta-analyses, as well as between homogeneous samples drawn homogeneous from populations and heterogeneous *metasamples* drawn from heterogeneous *metapopulations*. In so doing, we will show how "Contents" and "Effects" meta-analyses are designed to be complementary and how the former functions to address some of the concerns previously identified within the latter. Finally, to illustrate how one may use these complementary methods in conjunction, we direct readers to the second paper in this series (Figueredo, Black, & Scott, this issue), which demonstrates the utility of Contents Meta-Analysis with an empirical

example and presents findings regarding the generalizability of the effect sizes estimated.

Traditionally, the main forum for discussing empirical and theoretical discrepancies has been in the introduction section of academic papers, in which authors present arguments for and against the various points of view and draw a conclusion of their own. In theory, these introductions should be written based on a thorough review of the literature so that all appropriate primary sources are included. This rarely, if ever, happens in practice due to the sheer volume of available literature and the lateral spread (i.e., the diversity of research approaches and applications within or across research areas) of its content. Another major limitation of this approach is that the analyses done by the reviewers are *qualitative* in nature. As objective as we may try to be, unintentional biases in our selection and presentation of the supporting literature are almost certain to occur. We do this, knowing that the various precautions and controls that serve as the very foundations of the scientific method are designed to minimize the risk of bias at all levels of the research process. Why, when we generally take such pains to reduce sources of error using methodology and statistics, do we not always take the same measures in the process of research synthesis?

Meta-analyses address this limitation by offering a procedure to quantitatively analyze the literature with the objective of estimating a mean effect size for the population of extant scientific studies and then examine how the presence or absence of certain study characteristics might moderate effect size variation. We propose a complementary meta-analytic technique, termed *Contents* Meta-Analysis, as a supplementary analysis that permits investigators to produce a quantitative, evidence-based taxonomy of the available types of studies, identify clusters or patterns in methodological strategies, and employ these factors as predictive constructs in structural models.

The challenges of research synthesis and their relationships to scientific progress have been addressed in some fashion since the early 1900s (see Cooper & Hedges, 1994, Ch. 1 for a review). However, the work that invited a more thorough, albeit sometimes contentious, discussion of a systematic, *quantitative* method of research synthesis occurred in the 1970s with Tom Cook in 1974 and Gene Glass in 1976. They argued that despite the familiarity with and standard of using statistical analyses on primary data, the techniques could be modified such that they could just as well be conducted on secondary data. The advantages of practicing quantitative analyses of secondary data are numerous, not the least of which is that it provides a systematic way to analyze, and possibly even synthesize, discrepant results in the literature.

Still, the field of meta-analysis is not without its critics. Secondary data analyses preclude the use of the experimental method. Meta-analysts are

only able to observe what already exists in the literature, so manipulation of variables is impossible. Consequently, practitioners of meta-analysis must take the same precautions to assure internal and external validity that are necessary for the design and implementation of primary data studies. For example, sampling errors in meta-analysis include selection bias and non-sampling error, but techniques have been developed to minimize their effects. During data collection, a variety of sources may be used to review and extract relevant research, including both published and unpublished work. This strategy minimizes publication bias and concerns about the rigor or "quality" of unpublished works can be tested empirically using moderator analyses. Publication bias may also be identified using funnel plots that provide a graphic representation of missing data and can be quantitatively evaluated using a Pearson correlation coefficient. Additionally, fail-safe numbers estimate how many null studies outside of your sampling pool would need to exist to produce an overall effect size that is statistically nonsignificant (Card, 2011).

Methodological Problems: The Limitations of Conventional Meta-Analysis

The Problem of Heterogeneity

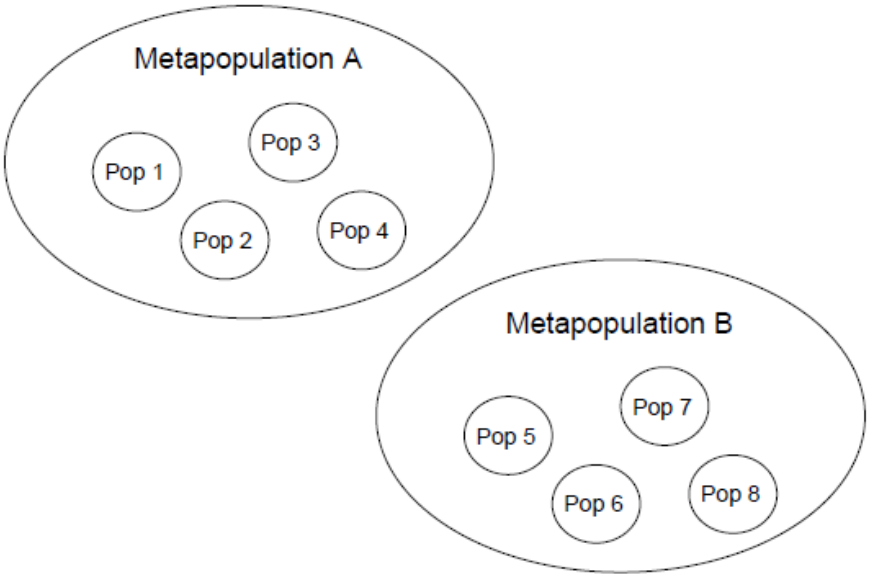
The standard function of meta-analysis is to estimate either a population parameter by averaging effect sizes across studies. Typically, meta-analysts will be aggregating data for a single bivariate relationship across a sample of studies. Interpreting this initial mean effect size depends on whether the effect sizes of the sample are homogeneous. In other words, when the effect sizes are homogeneous, they are all randomly sampling the *same* unitary population effect size. If the test for homogeneity rejects the null hypothesis, this indicates that the effect sizes are instead estimating systematically *different* population parameters, and that previously unspecified "follow-up analyses" are necessary.

The nonexperimental nature of meta-analysis leads critics to point out that methodological heterogeneity precludes one of the fundamental goals of meta-analysis, which is to produce an average effect size for the "population" of studies that one is sampling. They argue that legitimately pooling results across studies would require identical study characteristics in the sample to be synthesized. Although meta-analytic techniques have been developed to take advantage of heterogeneity across studies, variation in operational definitions, measurement, design, and methods may hinder the meta-analyst from combining outcomes into a single statistic.

To achieve greater clarity of exposition, we therefore propose that a terminological distinction be introduced to distinguish between a mean

effect size that is estimated for a *homogeneous* population of studies and one that is estimated for a *heterogeneous* population of studies. Unfortunately, the current literature continues to refer to the synthetic estimate as a "population" effect size even when it is found to be heterogeneous and therefore, by definition, to not describe the parameters of a *single* population but to instead describe an aggregate of *several* component populations, each with systematically different mean effect sizes. To sort out this terminological ambiguity, we propose to borrow the term *metapopulation* from biology for the heterogeneous case and reserve the term *population* for the homogeneous case (Levins, 1970). Levins (1969) defined a *metapopulation* as a "population of populations", consisting of several parametrically distinct populations that are localized within it (see Figure 1 for a graphical representation). Similarly, just as a sample of studies can be used to estimate the mean effect size for a single, homogeneous population, a *metasample* of studies can be used to estimate the mean effect sizes (plural) for the heterogeneous set of mutually discriminable populations comprising the metapopulation. If data are synthesized across this heterogeneous assemblage of local populations, as well as merely within them (as is often recommended), then the results of data aggregation from the entire metasample of studies can be used to estimate the *metapopulation mean effect size*, as long as the magnitude of the dispersion around this central tendency is also specified. As will be

Figure 1. A Schematic Diagram of Populations Nested Within Metapopulations.



described below, this can be accomplished by means of Generalizability Theory (GT) analyses, as adapted to meta-analytic applications. The relation here is analogous to that between the "group mean" and the "grand mean" in traditional analyses of variance, which can be estimated and have substantive importance even if the group means are significantly different.

An advantage of synthesizing studies with heterogeneous characteristics is that it provides quantitative compensation for the bias associated with any given particular study characteristic. Eluding the skewing effects of bias and its diversion of the trajectory of science away from the "golden fleece" of truth is a cornerstone of the scientific method. Multiplism and the auxiliary tools for selection among methodological options have long been strategies promoted by historians and philosophers of science to address the problem of research bias. Chamberlin (1890) stressed that active precautions be taken to minimize bias in the development and testing of hypotheses. His proposed strategy encouraged scientists to adopt the method of multiple working hypotheses, reasoning that the diversification of investment into several theories diminishes the likelihood of a special affinity for any one hypothesis that might lead to confirmation bias. Correspondingly, Platt's (1964) method of strong inference provided a guide to facilitate selection among the hypotheses following the Popperian canon of falsifiability.

The Sources of Heterogeneity

The observed effect size derived from any individual study is limited in terms of generalizability because the causal relationship under investigation is accompanied by variance from other sources, such as the testing method, setting, time, or any number of other variables. Cook (1993) termed these study components "irrelevancies" as they are systematically produced but not directly related to the causal relationship of theoretical interest. For example, if a construct such as optimism was measured using the Rainbows and Unicorns Scale (Extended)™, an individual's score would consist of (at a minimum) the "true" optimism score plus "test-specific" variance associated with the measure itself. By measuring optimism using several convergent methods, one may kill two birds with one stone: acquire a value that better approximates the true optimism score, and generalize across conditions that constitute the so-called *Heterogeneity of Irrelevancies*. As an added bonus, we may treat the irrelevancies as predictors in a model to produce more nuanced descriptions of optimism under different conditions. Of course, using them as a systematic source of variance disqualifies them as true "irrelevancies".

Nevertheless, some of these so-called "irrelevancies" may constitute the major threats to validity of meta-analytic models. Matt and Cook (2009)

provide a thorough outline of threats to the validity of meta-analytic models, including problems with sampling bias, underrepresentation of key attributes, coding methodology, and rater drift, among many others. For example, with regard to sampling bias, as in primary research, random sampling of the unit of study is very uncommon. The consequence of this is that sampling error is not randomly distributed so characteristics may be inadvertently (but systematically) weighted more heavily in certain groups. A related problem involves the inclusion criteria used to select studies for analysis, where those studies that were excluded may differ systematically from the sample pool, limiting inference from the sample to the target population.

Although the sources of threats to validity described by Matt and Cook (2009) are important, there are additional sources that are commonly left unaddressed in meta-analyses. For instance, close review of study design features may reveal systematic associations between study characteristics, or combinations of study characteristics, and research outcomes. Campbell (1986) used the term *local molar conditions* to highlight the importance of qualifying one's conception of internal validity as one that is nested within a particular set of circumstances rather than being a pure estimation as the theoretical construct implies. This conceptualization serves as an alternative to the "heterogeneity of irrelevancies", although it describes similar phenomena, but characterizes them as potentially worthy of theoretical interest.

In other words, even a true experimental design is not so pure in the Platonic sense; we are never truly reaching the Platonic *εἴδωλον* (transcendent ideal) of experimental design because any individual manifestation is a product of the experimenter who is still acting on his or her own limited experience. Thus, the term "local" is used as a qualifier to recognize that we are also sampling the temporal and spatial context in any given study. Both the setting and time in which a study is conducted are conditions under which an effect may or may not occur. The term "molar" refers to the kinds of discrete "packages" of methods or treatments tested by researchers that have produced apparently meaningful results and therefore continue to be used by the scientific community at large. Its use is also intended to acknowledge the fact that a given experimental method or treatment is *constructed*, as well as to grant it the right to be evaluated as it is, rather than against a pure, theoretical standard.

Kuhn (1970) argued that the content and methods of science, and therefore its conclusions, depend on the current research paradigm. Each paradigm is encapsulated, operating under different assumptions, and using different standards for evidence. As such, the methodological variables that are available for investigation will correspond to the dominant paradigm, and will potentially change when a new paradigm takes over. Even if some remnants of prior methodological trends survived

a paradigm shift, by Kuhn's definition, outcomes of the two paradigms would be incommensurable. Following Kuhn's proposal, several others proposed modifications in order to rectify perceived limitations in his original model. Feyerabend (1975) took a completely different approach to defining paradigms and ideologically opposed the concept entirely. His view encouraged independent thought and eliminating the influence of ruling institutions and requirement to be consistent with others. He reconstituted Kuhnian paradigms as completely socially constructed, devoid of rationality and only reflecting ideas of society. Feyerabend agreed that paradigms create their own standards of evidence and this characteristic makes them resistant to change, but went beyond this assertion to argue that even the "context of justification" is a paradigmatic perspective.

Lakatos (1978) introduced the idea that several research programs compete simultaneously. Laudan (1977) also disagreed with Kuhn's characterization of paradigms as isolated and independent phenomena. Instead of long periods of stable dominance of a single paradigm, punctuated by rapid shifts to a new paradigm once the first paradigm became untenable, as Kuhn originally proposed, both Lakatos and Laudan argued that science evolves gradually. In Laudan's "Research Traditions", specific facets or aspects of a paradigm could be changed as necessary and all were potentially replaceable without changing the underlying tradition. Under this view, the criterion for accepting a Research Tradition was its effectiveness in solving problems. In line with Laudan, Lakatos proposed a slower process of change, but eliminated the serial element common to both Kuhn and Laudan. His model of scientific progress described the differences between what he termed *Progressive* and *Degenerative* "Research Programs". Whereas Progressive ones explain more phenomena parsimoniously without increasing in complexity, Degenerative ones fail to make new predictions and become more elaborate to account for contradicting phenomena.

This brief (and probably oversimplified) foray into the field of metascience is intended to drawing attention to the fact that any given research study is nested within a particular *Zeitgeist* (whether one calls it a Research Paradigm, Tradition, or Program). Critics may rightly respond that the influence of a particular *Zeitgeist* on research is perhaps one of the most difficult, if not impossible, sources of dependence to account for, at least in the Kuhnian sense of the term. As Feyerabend noted, even the context of justification exists within a socially-constructed framework and is thus paradigm-bound. We cannot observe the paradigm directly, just as an astronomer or physicist cannot observe the universe directly. We are only capable of trying to understand its structure through our limited methods of measurement. However, it is possible to reframe one's conception of a paradigm to a more manageable construct that can be

clearly identified and delineated from other constructs. For instance, one may consider certain academic disciplines to possess different research paradigms. In the social sciences, psychology, sociology, anthropology, and economics all presumably adhere to different paradigms. In psychology, the field may be further disaggregated into social, clinical, or neuroscientific fields. In framing paradigms this way, it is important to recognize that these constructs may be nested within larger constructs, just as the researcher is nested within a particular laboratory. Once a paradigm has been operationalized, either through a priori research design planning or using exploratory measures, researchers conducting Contents Meta-analyses may quantitatively compare the effects of those research contexts on study outcomes.

Statistical interdependence among study features therefore results in quantitative bias and obfuscates correct interpretations of the data. Although the overall fit of the model of interest may not be significantly affected by interdependence of predictors, multicollinearity will certainly change the estimations of individual parameters by inflating their standard errors and biasing our estimates of their effects. This problem is undesirable in any case, but especially so when testing structural models in order to make claims about causality. Predictors that may be causal in reality are rendered seemingly inert under conditions of multicollinearity. This potential to commit a Type II error in subsequent models without the affected predictor will produce incorrect model estimates, further contributing to uncertainty or ambiguity of causal influence, constituting a threat to the internal validity of the structural model. Researchers therefore need to account for local molar conditions in their meta-analytic studies in order to avoid distorting results.

Figueredo (1993) discussed this problem, noting that these threats to generalizability in research synthesis cascade out of a single, common obstacle in meta-analyses: *violations of independence* among studies in the metasample. As any undergraduate with a basic understanding of ANOVA can tell you, the validity of the conclusions drawn depend on whether certain analytical assumptions are met; one of those assumptions is that of mutual independence among observations. When this assumption is violated, and data from dependent samples are synthesized and interpreted as if they were from independent samples, the mean effect size would be skewed in the direction of the results of those studies. Unfortunately, the results of any meta-analysis may be more susceptible to this violation for less obvious reasons than those articulated by previous researchers who focus on dependence within studies. A hierarchical taxonomy of understanding these methodological dependencies among studies may thus be constructed as follows:

1. *Multiple Studies by Single Researcher*. The first source of dependence among studies lies with the single researcher. A typical

academic scientist conducts his or her research program by testing several related hypotheses to gain ground in a particular field of study. The way that research is conducted depends on a number of factors, such as funding, institutional support, previous experience, graduate studies, and the proclivities of his or her graduate advisor. Those characteristics represent a sample of possible characteristics held by someone conducting scientific research. Therefore the methods and conclusions produced by that researcher are limited by those parameters. A meta-analyst who includes several studies by the same researcher, then, will inadvertently bias the sample in the direction of the characteristics of that researcher.

2. *Multiple Researchers from Single "Laboratory"*. Similarly, the intellectual context in which a single researcher operates is partially a function of the scholars in the immediate vicinity, all of whom come to the table with a particular set of skills and knowledge. Graduate students in the Ethology and Evolutionary Psychology program at the University of Arizona, for example, produce research in variable topics including attachment theory, morality, psychopathy, facial expressions, cross-cultural studies, behavior genetics, female fertility and mating, spatial distributions of different populations and so on. Regardless of this topical diversity, the underpinnings of those works are remarkably similar in their essence, clearly reflecting the constellation of characteristics that make up the research strategies, approaches, and areas of interest of our graduate advisors and their programs of research. In meta-analysis, the problem remains that some perspectives may be oversampled while others are not.
3. *Multiple "Laboratories" within Single Research Paradigm*. Laboratories themselves are also nested within a set of conditions that may influence dependence in meta-analytic studies. For instance, *temporal* conditions play a role in the types of technology available with which one may conduct research, or in methodological or theoretical advances made in the field upon which research may be based. *Political* conditions may dictate what areas of research are being funded or perhaps even how "academic freedom" is treated (but hopefully not). *Intellectual* conditions influence the major theories that are in vogue, which have the potential to shape research fundamentally, in terms of the questions, variables, methods, and interpretations generated by investigators. These socio-cultural ecological factors, or *research paradigms* (see Kuhn, 1970), have the potential to influence research outcomes, but are nevertheless neglected in traditional meta-analytic studies.

These dependencies among the data present a problem for those who view meta-analysis as a culmination of methodological advances designed to address the epistemological limitations of traditional methodologies that were brought to our attention by our philosophical forebears. Indeed, the fundamental structure of meta-analytic methods, that of acknowledging and incorporating multiplism in the research process, pays tribute to Chamberlin's (1890) multiple working hypotheses, Platt's (1964) strong inference, and Cook's (1985) and Shadish's (1993) critical multiplism. Still, the field may benefit from a review of the claims of philosophers and methodologists about how and when progress in science occurs, and what that may mean for researchers conducting meta-analyses.

Methodological Solutions: Two Complementary Techniques of Meta-Analysis

Given the limitations of conventional meta-analysis, Figueredo & Scott (1992) proposed a complementary method to address some of the threats to internal and external validity described above. Applying this technique would not demand a significant cost of time or effort over and above what is already expended for a typical meta-analysis, but it directly addresses violations of independence of meta-analytic observations, such as similar design features across studies.

The Strategy of Critical Multiplism

The logic of multiplism was first extended to other aspects of research, in addition to the formulation and testing of hypotheses, when Campbell and Fiske (1959) proposed multioperationalism in psychometric measurement and demonstrated its utility for construct validation using the Multi-Trait Multi-Method Matrix (MTMM). MTMM analysis enabled researchers to quantitatively disaggregate trait variance from method variance and measure convergent and divergent validity. Attendant techniques for selecting among possible MTMM models using confirmatory factor analytic methods were later introduced by Widaman (1985). He presented a procedure with which to specify latent variable models for use with MTMM data and to use hierarchical nested model comparisons to produce more precise estimates of trait variance and method variance, and to test the degree of convergent and divergent validity. These combined contributions equipped researchers with a quantitative solution to account for bias associated with different methods of measurement.

Shortly thereafter, the multiplist movement evolved to encompass all aspects of the research process with *critical multiplism* (Shadish, 1993).

This fully inclusive model of multiplism in methodology expanded from the previous applications to hypotheses and measurement to the selection of theoretical frameworks and models, research designs, methodologies, statistical analyses, interpreting results, and summarizing literature. Although multiplism at this level of complexity was not explicitly defined until the late 20th century, John Stuart Mill published his methods for inferring causality, one of which alludes to strategic use of multiplicity in methods, as early as 1843. The Joint Method of Agreement and Difference prescribes a symbiotic relationship between naturalistic methods that operate using the Method of Agreement, and experimental methods that operate using the Method of Difference. Mill's method and the principles of multiplism were applied in a recent enterprise by social psychologists Mortensen and Cialdini (2009). Termed "full-cycle" social psychology, the authors recapitulated the biases inherent in laboratory and naturalistic studies but noted that, when paired, both the abilities and limitations of each method were complementary. Thus, if implemented in a cyclical fashion, where naturalistic observations are followed by review of current theory to explain the phenomena in question, empirical tests are conducted to test hypotheses predicted by theory, and the experimental results are then corroborated in a naturalistic setting, the limitations associated with any one method are minimized.

Fundamental to all multiplist approaches to research is the idea that different types of methodology constitute systematic sources of bias. Shadish (1993) noted in his technical guidelines that, when employing critical multiplism, researchers should "note" (p.20; with no mention of a quantitative method) any moderating effects of particular methods and account for differences in results associated with methods with different biases. However, a comprehensive application of critical multiplism would demand virtual omniscience (which would preclude the need for scientific research in any case) to know all tasks entailed by a particular research question, the options for implementing those tasks, and their associated biases. Thus, Shadish prescribes enlisting the help of sources, including people and competing theories, whose biases differ from those of the primary investigator.

The "critical" aspect of this proposed methodology must be emphasized. It refers to an attempt, by empirical or analytical methods, to identify systematic bias associated with different research options. It can be contrasted with "mindless" multiplism, which is a (clearly inferior) way of implementing multiplism without thinking about the contributions or costs of the options chosen (Shadish, 1993). As noted above, the use of a limited version of critical multiplism by an individual researcher is possible and would indeed strengthen any resulting inferences. An alternative, more efficient implementation (although in no way meant to dissuade individual researchers from integrating multiplism into their

work) has been advocated by Shadish and others (e.g., Campbell, 1987; Figueredo, 1993) and entails adopting critical multiplism at the institutional level. Under this model, researchers in a given area of study, with unique and complementary skills and knowledge, would become contributors to addressing research problems systematically and systemically. Collaborations would take a new form, with diversity as the defining feature, rather than the status quo where "birds of a feather flock together". A major institutional shift would be required in the sciences. Indeed, there are increasingly new initiatives to provide transparency in research, make data sets public, and provide a forum for work that would otherwise be left in the "file drawer". In a number of ways, however, the current academic climate, and especially its incentive structure, may not have the requisite elements in place to foster a comprehensive application of critical multiplism.

In the meantime, what can be done? As Figueredo (1993) pointed out, meta-analyses can easily accommodate multiplism; indeed it is woven into the very fabric of this type of analysis. Moreover, the multiplism supported by meta-analytic techniques can be critical, per the definition offered by Shadish. In fact, it may be the ideal tool for a critical approach because it can provide a quantitative analysis of the various options present across studies. As with Widaman's extension of MTMM analysis to a more precise quantitative test of method variance, a correspondingly advantageous tool to estimate the contribution of all variations of all of the research elements to research outcomes is invaluable. Of course, since this process is retroactive in nature, there will always be a finite number of variations for each research element. At the aggregate level, it is also reasonable to suppose (and stands to be tested) that any given methodological or procedural variation will be reported multiple times in the literature, due to the predominant research paradigm or available technology.

Critical multiplism implicitly operates under the variance components model where a given score in a distribution represents the amalgamation of the "true" score, plus systematic variance associated with methodological characteristics, plus unsystematic variance associated with apparent stochasticity. Meta-analysis provides a vehicle with which we can tease apart these elements quantitatively. Once we calculate values from aggregate data that describe study characteristics, it will be possible to identify correlated design features and clusters of research techniques which can then be analyzed using more complex statistical techniques such as exploratory and confirmatory factor analyses, and structural equation modeling.

The implications of having these types of tools at hand are non-trivial. For instance, the analysis of secondary data shares the limitation associated with primary data research with regard to the degree of generalizability based on inclusion criteria. Meta-analysts face the

daunting task of specifying the exact criteria that study inclusion will be based on. Decisions must be made in the service of the research hypotheses being tested and, more practically speaking, with the consideration of the available budget, timeline, and personnel. Key determinations for study inclusion are the operationalization of variables, sample characteristics such as demographics and diagnostic criteria, the methods of measurement, study design, the specific regions from which data will be sampled, the time frame for data retrieval, and the types of publications that will be included, ranging from refereed journal articles to dissertations and conference proceedings and even possibly unpublished work left for an ignoble death in the infamous "file drawer" (Card, 2011). The thus far untapped advantage of meta-analysis, however, is that it provides an additional opportunity to quantitatively measure methodological limitations and exclusions in primary research at an aggregate level. Once they have been identified, it is then possible to *prescribe* future research directions that can rectify bias in methodological paradigms. One broader implication is the impact on the generalizability of meta-analytic results, which would be substantially enhanced with a body of science *proactively* employing critical multiplism (Cordray, 1986)

Pooling the Results

Once data have been extracted from the sample of studies and converted to an appropriate common metric of effect size, we have the data pool upon which all subsequent analyses will be conducted. We begin by generating descriptive statistics about the sample, with the key difference being the unit of analysis. In the primary data analyses, we describe the average person using measures of central tendency such as the mean, median, or mode. Similarly, we can calculate those measures for the sample of effect sizes and often the statistic of interest is the mean.

A mean effect size is computed by dividing the sum of the effect sizes multiplied by its corresponding study "weight" by the sum of the study weights, where study weight may be a function of the standard error (e.g., $1/SE^2$). The standard error for the mean effect size should also be calculated so that the resulting statistics can be evaluated for significance by calculating a z statistic (the mean effect size divided by the mean effect size standard error). You may note the irony in this resulting step, which is in direct contrast to our earlier criticisms of null-hypothesis statistical testing (NHST). We find it more than a little odd that some of the same methodologists that criticize the utility of NHST nonetheless, especially when overpowered (as by using larger sample sizes), tout the benefits of meta-analysis in increasing the power of NHST (as by pooling samples sizes so as to make them collectively larger).

Testing for Homogeneity

Thus, although we begin by avoiding NHST, we end up by using it anyway. Nevertheless, this application of NHST is not without some function. Rather, it is just a first step leading to a series of options with which we can use to analyze these effect sizes. The next step is quite important, although it almost always produces the same result, and that is the evaluation of the homogeneity of the sample.

The sampling distribution of effect sizes can reveal whether sources of variability are limited to sampling error, or whether sampling error is "supplemented" with other variability arising from study differences. The variability among mean effect sizes in the metasample, in comparison with the theoretically-constructed "standard error of the mean" that can be estimated under the "null hypothesis" that they are all drawn from a single homogeneous population, is our first indication that the data are not homogeneous. The estimation of this additional variance component permits us to estimate the dispersion of the individual parameters of the *component* populations about the aggregate parameter that can be estimated for the central tendency of the metapopulation as a whole.

However, if we were not willing to accept subjective inferences about data synthesis before, we are certainly not going to start now. A quantitative, inferential test of homogeneity may be conducted which produces an estimate of whether all effect sizes could have been drawn randomly from the same unitary population parameter under the Central Limit Theorem. This Q-statistic is distributed as a Chi-square with $(k-1)$ degrees of freedom, where k is the number of studies. The resulting value may be tested, again using NHST, and it almost always produces a significant result, indicating that the metasample is indeed heterogeneous. In other words, at least one effect size parameter in the metasample is derived from a constituent population with a systematically different mean than that of the metapopulation as a whole.

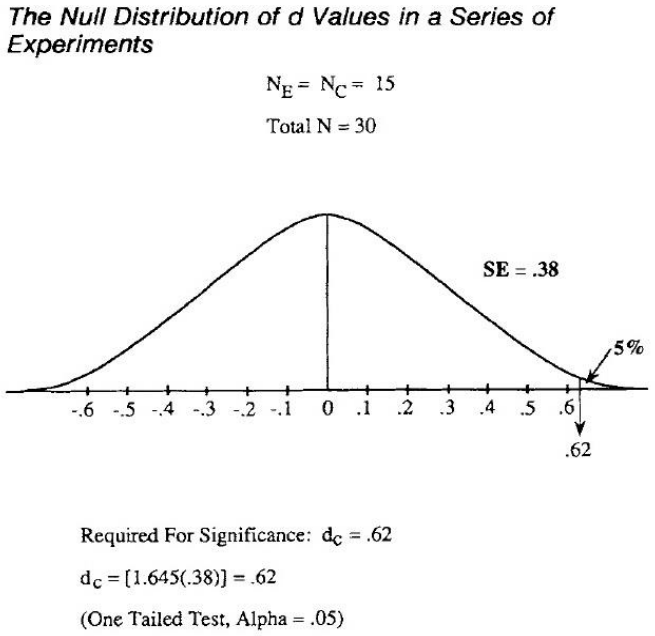
Enhancing Statistical Power

Contemporary meta-analytic methods (e.g., Hedges & Olkin, 1985) are lauded for their use of the effect size, rather than the significance test, as the unit of analysis (although some older meta-analytic methods did involve significance tests). The lackluster performance of the p-value as a useful criterion derives from its limited informational content, only indicating whether the effect in question is statistically non-zero. Additionally, as Rosenthal and DiMatteo (2001) and many others have pointed out, the result of any given significance test is a function of the effect size and of the sample size (p. 63). Thus, with a large enough sample size, a variable with even a small effect size may be found to be *statistically*

significant, which only indicates that the effect size is "not statistically equivalent" to a population parameter of zero.

Schmidt (1992) argued that conclusions based on hypothesis testing are fundamentally misleading and showed that pooling effect size data has an added advantage of increasing statistical power[1]. He provided a hypothetical example wherein the *true* effect size in the population for a drug is .50. In the null distribution of this example, the mean is zero, with a standard error of .38. In both distributions, variation about the mean is due to sampling error. Using a one-tailed test with alpha equal to .05 would require the observed effect size to be .62 or larger to identify a significant difference. With no effect in the population, only 5% of observed effect sizes would meet or exceed that value (see Figure 2).

Figure 2. The Null Distribution of d Values in a Series of Experiments. Reproduced with permission from Schmidt (1992).

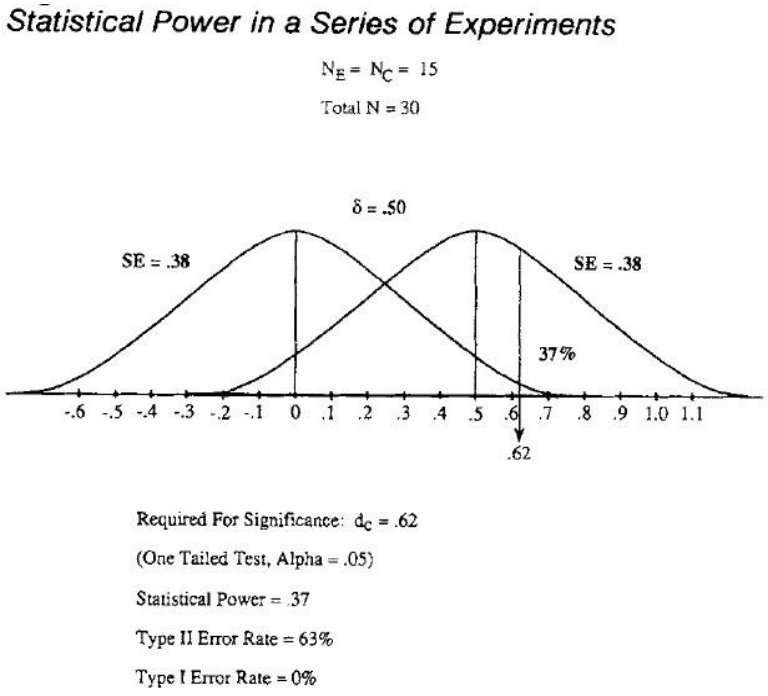


However, the true population effect size is .50, as previously noted. This means that Type I error is actually zero because it is impossible when an effect actually exists in the population, and only Type II error can occur. It also means that obtaining an effect size of .62 or larger will only occur in 37% of studies conducted (see Figure 3). A staggering 63% of studies conducted would lead to a false conclusion. Furthermore, estimates of the population effect size would be distorted because the mean effect size

[1] The relationship between meta-analyses and statistical power will be described in further detail later in this paper.

estimated from significant tests of the hypothesis is well above the true population effect size. At a minimum, the lowest effect size associated with a "significant" result is 24% above the true value of .50. It is quite clear that there is a problem with significance testing when an observation equal to the true population effect size would lead to the conclusion that there is no effect. Schmidt then shows that a meta-analysis yields the correct conclusion. The average effect size will approach the population effect size if the number of studies is large and any sampling error will average to zero.

Figure 3. Statistical Power in a Series of Experiments. Reproduced with permission from Schmidt (1992).



This brief example by Schmidt (1992) illustrates how meta-analysis can compensate for the low statistical power typically found in individual studies in the behavioral sciences (Cohen, 1962, as cited by Cohn & Becker, 2003). Statistical power is quantitative and measures the probability of detecting an effect that truly exists. Unfortunately, as Cohn and Becker point out, traditional narrative syntheses of scientific literature do not account for the low power available in some studies, and this problem can create the illusion that an effect does not exist when it actually does. Although it is often said that enhanced statistical power results from the larger sample size as results from the sample of meta-analytic studies are

pooled, Cohn and Becker argue that this inference is incorrect. We now explore some of the reasons why this might be true.

Frequently Rejecting Homogeneity

In all likelihood, the sample of studies in a meta-analysis will include methodological characteristics that vary from study to study. Methodological variation is directly related to variation among the parameters of different populations represented in the metasample, so the assumption of homogeneity is usually rejected in meta-analyses (Osburn & Callender, 1992 as cited by Hunter & Schmidt, 2002).

This insight is a mere stepping stone to a world of possibility where various analyses may be conducted, but it is immediately clear that we are not finished with our work and it would be inappropriate to simply report the mean effect size from our original synthesis. Possibilities include looking for moderator variables, as one might do in traditional analyses in an ANOVA or regression analysis, or accounting for between-study variance in a random-effects model and using that information to synthesize an adjusted metapopulation mean effect size for the aggregated heterogeneous sample.

The choice of follow-up analyses depends on a number of factors to be evaluated by the investigator, but the end goal is the same: to determine what knowledge may be gleaned from a quantitative synthesis of data. Central to that goal is the estimation of the metapopulation effect size, properly adjusted for or qualified by systematic between-study differences. The quantitative method of data aggregation, in our estimation, is inherently superior to the more frequently employed, although less burdensome, standard literature review.

Performing Causal Analysis of Discrepant Results

There are a number of possible follow-up analyses which one may use to go further and *model* the variance among effect sizes in a heterogeneous metasample. Such a model is intended to provide a *causal explanation* of the observed heterogeneities. One possibility is to conduct moderator analyses, wherein sample or methodological characteristics are coded as independent variables in an ANOVA, and another is to use these sample or methodological characteristics as predictor variables in a multiple regression analysis that predicts the effect size as the criterion variable. More sophisticated analyses of these systematic effects may involve structural equations modeling and generalizability theory analyses, which may also use a combination of random and fixed effects.

Fixed-effect models assume effect sizes are homogeneous and estimate a single population parameter. These models are limited in their

informative content and generalizability. A common analogy is with the analysis of variance (ANOVA), where the levels included in a model are supposed to represent all possible levels; thus, with regard to generalizability, a study would have to share the same study characteristics as those included in the fixed-effects model to draw any conclusions. They have also been criticized for overstating precision and distorting conclusions (Hunter & Schmidt, 2002). This is because the sample size of a study influences the overall population estimate, which is assumed to be the best estimate of the population effect size in fixed-effect models. Specifically, a study with a larger sample size will produce an estimated effect size with a smaller variance than a study with a smaller sample size. These estimates are more precise and weighted more heavily in the calculation of the mean effect size, with the additional consequence that the confidence interval will also be smaller. As the confidence interval decreases, statistical power appears to increase, so each additional study included in the meta-analysis leads to the appearance of higher statistical power (Cohn & Becker).

In contrast, a random-effect model assumes heterogeneity of effect sizes, which indicates that at least one effect size has been sampled from populations with different mean effect sizes. In other words, a random-effect model is used when population parameters are not the same in all studies in the *metasample*, so it assumes effect sizes are sampled from a *metapopulation* distribution which is greater in dispersion than that which would be expected from mere "sampling error" around the central tendency of any constituent population, consequently providing an estimate for the mean and variance of that heterogeneous metapopulation. Heterogeneity is quantified by calculating tau, and the mean effect size and standard error are adjusted accordingly. As additional studies are included in the meta-analysis, each with their own associated population parameters (according to the more localized component population from which each study in the metasample is drawn), the tau statistic may fluctuate and this leads to a potential increase in the estimated standard error, which in turn appears to decrease statistical power (Cohn & Becker). Cohn & Becker (2003) therefore distinguish the effects of data pooling on statistical power within fixed-effects versus random-effects models.

A *random effects* model therefore represents a useful technique to model the heterogeneity among effect sizes because this method requires estimating the metapopulation variability in effect sizes, which is a function of: (1) the heterogeneity of the metasample, (2) the number of studies analyzed, and (3) the weight assigned to each study. The *tau* statistic, which estimates between-study variance, and an estimate of the sampling variance are combined to create a new weight that will be used in calculating the mean effect size and standard error for the heterogeneous

metasample (see Card, 2011). However, this method simply *describes* the heterogeneity and does not *explain* it.

An investigator may then proceed to examine the contribution of varying study characteristics to the variation around the mean effect size when that variation cannot be explained by sampling error alone. These "effects" meta-analyses are useful in terms of their ability to explain variation in effect sizes using methodological variables as predictors. Depending on research objectives, fixed-effects and random-effects analyses permit researchers to look at trends over time, or select a methodological variable that potentially moderates the relationship under investigation. Nevertheless, a meta-analyst is ultimately limited to the set of variables available in primary research, which represent only a subset of the total population of characteristics available for study.

This insight is not especially novel to scholars familiar with meta-analytic techniques, but we would like to extend the usual criticisms to encompass limitations associated with the current, dominant research paradigms. For example, consider an ambitious investigator who is interested in aggression or deviant behavior over the last 60 years. The population of available research spanning this period is subject to several paradigmatic shifts in psychology, including behaviorism and the subsequent "cognitive revolution". We argue that a diligent researcher would be remiss to not at least consider these contextual influences.

These sources of dependence in the primary data threaten the utility of the meta-analytic method as a tool to estimate generalizability. The method proposed here may serve the dual purpose of accounting for statistical dependence and serving as a compass for future research endeavors. Cordray (1986) and Shadish (1986) have both suggested that meta-analysis may be employed to assess gaps in the literature for the purpose of strategically planning subsequent research. In this approach, conducting research becomes a collectivistic undertaking and the merit of a study depends on its additive value to the entire scientific enterprise rather than solely on the creative novelty of an individual researcher.

Documenting the Design Features of Studies

The heterogeneity of effect sizes in a metasample may be a function of variation in methodological variables (Osburn & Callender, 1992 as cited by Hunter & Schmidt, 2002). This relationship can be modeled using structural modeling ("model-driven meta-analysis", Becker, 2009), which can provide a test of hypothesized causal relationships between different study characteristics and systematic variation in effect sizes. To be able to develop such a causal model of potential heterogeneities among observed effect sizes, it is necessary to first be in possession of the data that would be necessary to provide a basis for such an explanation. Thus, in the study-

coding phase of meta-analysis, researchers usually take note of methodological and procedural characteristics of the studies in their sample. The type and number of variables depend on theoretical and practical grounds, but Stock (1994) proposes a basic classification system that includes characteristics of the report, setting or study context, subjects, methodology, treatment, process, and effect size. In the more mature phases of meta-analysis, variables classified during coding are then used to find patterns in the relationships between study characteristics and study outcomes. For instance, analyzing effect sizes as a function of the year the study was published may inform us about trends over time.

For the purpose of looking at more complex phenomena than the standard single bivariate meta-analysis permits, Becker (2009) proposed the construction of an average correlation matrix. In this case, multiple bivariate relationships are of interest as well as the relationships among them. In brief, this method involves creating a correlation matrix for each study that includes all effect sizes of interest and then estimating the average effect size for each bivariate relationship in the matrix using a fixed- or random-effects model (depending on the characteristics of the sample and the goals of the investigator). The result is a synthesized correlation matrix upon which regression or path models can be imposed. This advanced methodology has the advantage of allowing researchers to examine indirect effects and mediating effects (Becker, 2009; Figueredo, 1993).

A major appeal of meta-analysis is its potential for the generalizability of results because its design is able to overcome limitations frequently present in primary studies. Insufficient power to detect true effects, as mentioned above, is a common problem plaguing primary studies, but this pestilence appears to be in remission in meta-analytic studies (although see discussion above as well as Cohn & Becker for a review of caveats of this assumption). Additionally, whereas any single study may be limited in the characteristics of its sample, methods, or procedure, a meta-analysis combines a variety of study characteristics, broadening its extrapolative reach.

"Contents" Meta-Analysis

Originally proposed by Figueredo and Scott (1992), a *contents* meta-analysis may be performed in conjunction with the conventional *effects* meta-analysis. While the latter may be used to examine the causal analysis of discrepancies in reported effects, it is still limited by the dominant research paradigm(s). This supplementary analysis is suggested in direct response to those limitations.

Rather than focusing on the results of studies, contents meta-analysis is analogous to a “content analysis” of the text of a study. Content analysis

is a method used to systematically analyze the properties of large amounts of text (Holsti, 1969). The process begins by identifying the key linguistic characteristics of interest, which will then be used in a coding system for text analysis. For example, one may code the presence or absence of key words and use the resulting values to compare different bodies of text and make inferences about their relative meaning. More broadly, this process of identifying and measuring key characteristics is a central concept in program evaluation, market research, and trend analysis, among others. Within the context of a contents meta-analysis, this same basic process is applied to the content of the text in a sample of studies. In this case, a body of literature is coded according to a set of criteria determined by the objectives of the researcher.

Specifically, the objective of a contents meta-analysis is to focus on the methods of studies by producing a quantitative analysis of their relative frequency and of correlated design features. The underlying argument is that the research tactics employed by a particular researcher, or within a particular laboratory or research paradigm, are not independent. Moreover, these tactics form "discriminable constellations of related elements" (Figueredo, 1993) that can be identified using common statistical methods.

One such method is common factor analysis, wherein patterns of common associations would be extracted from the sample of study features. For example, one may wish to identify whether certain disciplines tend to employ certain methods over others. Do we observe systematic differences in measurement methods across disciplines? Are psychologists more likely to use self-report whereas anthropologists may use naturalistic observations? Do medical researchers employ clinical interviews while economists use secondary data from national samples? More importantly, how does the method influence the outcome? If certain methods produce inflated effect sizes relative to others, we need to account for that.

An appropriate application of factor analysis includes both exploratory and confirmatory analyses of correlated design features. Exploratory factor analyses produce factors based on statistical criteria; in this case the process is quite atheoretical but may serve as a starting point for subsequent analyses. A researcher then may elect to test whether the factors produced by the exploratory factor analysis are replicable on an independent sample using confirmatory factor analysis. Alternatively, a researcher may construct factors according to theory and test their model. The results of these factor analyses constitute the "multivariate operationalization of a research paradigm" (Figueredo, 1993) and permits development of measurement models of metascientific constructs, or paradigms represented by the research proclivities of an individual, a laboratory, a discipline, and so on.

Either way, the use of common factors rather than individual indicator variables as predictors has all of the well-documented benefits that are observed elsewhere in predictive models. As compared with single indicator variables, when common factors are used to stand in for an array of convergent indicators of the same constructs, the multiple operationalizations that they offer possess all of the following advantages: (1) increasing the *reliability* of measurement of the predictive constructs; (2) increasing the *validity* of measurement of the predictive constructs; (3) decreasing the absolute *number* of predictors, and hence the complexity of the model; and (4) decreasing the *collinearity* among model predictors.

Relating "Contents" to "Effects" in Meta-Analysis

Although the prospect of identifying measurement models of paradigmatic constructs is exciting enough, the potential for contents meta-analysis as an informative tool is not yet exhausted. Once the factors are identified and estimated, we may evaluate whether the effect sizes produced from one paradigm construct are different from those produced by other paradigm constructs. In essence, we may account for measurement effects in outcomes at the theoretical, rather than just empirical, level. The resulting latent constructs may serve as predictors in structural models. These "meta-analytic factor-analytic structural equation models" (Bentler, 1989; Scott, Figueredo, & Hendrix, 1992) provide a more sophisticated method of relating methodological contents of studies to magnitudes of effects reported.

Employing these latent "paradigm" factors as predictors in a meta-analytic model has at least two advantages. First, they serve as a data reduction method for model predictors by absorbing shared variances among different, but associated measures into a single construct. This prevents the superfluous inclusion of predictors that unnecessarily absorb degrees of freedom and risk overfitting of the model. A second, but related advantage involves the ability to control spurious relationships associated with statistical interdependence among predictors. In using this method it is possible to identify and control for systematic distortions in effect sizes (*method* variance) that result from common methodological practices. Contents meta-analysis permits the establishment of meta-analytic latent variable models to empirically test both the existence and the defining parameters of extant research paradigms.

Summary and Conclusions

We have tried to establish the methodological rationale and justification for the more widespread application of Contents Meta-

Analytic methods by presenting the following series of arguments: (1) an overview of the functions, merits, and limitations of meta-analysis from methodological and meta-scientific standpoints; (2) the introduction of some novel terminology, highlighting the distinction between complementary methods of Contents and Effects meta-analysis, as well as between heterogeneous metasamples and metapopulations from homogeneous samples and populations; and (3) the rationale for using Contents Meta-Analysis as a supplemental technique to precede and increase the effectiveness of traditional meta-analytic methods.

In response to the limitations of conventional meta-analyses described herein, we are therefore recommending a variant form of meta-analysis, which we call *Contents* Meta-Analysis, intended to be complementary to the traditional model, which we call *Effects* Meta-Analysis, to distinguish it from the former. Contents Meta-Analysis is designed to turn an erstwhile limitation, multicollinearity among predictors in a meta-analytic model, into a tool to explore the patterns that may exist in research practice that may compromise the non-independence of meta-analytic observations at the level of the studies sampled. We believe that the more widespread application of this method can not only enhance the practice of meta-analysis, based on the statistical problems that it solves, but also improve our understanding of the structure of the scientific literature that meta-analysis is meant to describe.

References

Becker, B. J. (2009). Model-based meta-analysis. In H. M. Cooper, L. V. Hedges, & J. Valentine (*Eds.*), *The handbook of research synthesis and meta-analysis. (2nd ed.)* (pp. 377-395). New York: Russell Sage Foundation.

Bentler, P. M. (1989). *EQS: Structural equations program manual*. Los Angeles: BMDP Statistical Software.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W.M.K. Trochim (*Ed.*), *Advances in quasi-experimental design and analysis. New Directions for Program Evaluation, no. 31*. San Francisco: Jossey-Bass.

Campbell, D. T. (1987). Evolutionary epistemology. In P. A. Schilpp (*Ed.*) *The philosophy of Karl Popper (*, pp. 413-463*),* LaSalle: Open Court.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Card, N. A. (2011). *Applied meta-analysis for social science research*. New York: The Guilford Press.

Carpenter, S. (2012). Psychology's bold initiative. *Science, 335*, 1558-1561.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science, 15*, 92.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*, 243-253.

Cordray, D. S. (1986). Quasi-experimental analysis: A mixture of methods and judgment. In W.M.K. Trochim (*Ed.*), *Advances in quasi-experimental design and analysis. New Directions for Program Evaluation, no. 31*. San Francisco: Jossey-Bass.

Cook, T. D. (1974). The potential and limitations of secondary evaluations. In M. W. Apple, M. J. Subkoviak, & H. S. Lufler (*Eds.*) *Educational evaluation: Analysis and responsibility*. Berkeley, CA: McCutchan Publishing.

Cook, T. D. (1985). Postpositivist critical multiplism. In L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21-62). Beverly Hills, CA: Sage

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relations. In L. B. Sechrest and A. G. Scott (*Eds.*), *Understanding causes and generalizing about them. New Directions for Program Evaluation, 57*. San Francisco: Jossey-Bass.

Cooper, H. M. & Hedges, L. V. (*Eds.*) (1993). *The handbook of research synthesis*. New York: The Russell Sage Foundation.

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London: Verso.

Figueredo, A. J. (1993). Critical multiplism, meta-analysis, and generalization: An integrative commentary. In L. B. Sechrest (*Ed.*) *Program evaluation: A pluralistic enterprise*. Special Issue, *New Directions for Program Evaluation, 60*. San Francisco: Jossey-Bass.

Figueredo, A. J., Black, C. J., & Scott, A. G. (this issue). Complementary meta-analytic methods for the quantitative review of research: An extended illustration.

Figueredo, A. J., & Scott, A. G. (1992). An example of contents meta-analysis and its use in higher education retention studies. Paper presented at the American Evaluation Association Annual Meeting, Seattle.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

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

Holsti, Ole R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

Hunter, J. E. & Schmidt, F. L. (2002). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275-292.

Kuhn, T. S. (1970). *The structure of scientific revolutions. (2nd ed.)* Chicago: University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programs*. Cambridge, England: Cambridge University Press.

Laudan, L. (1977). *Progress and its problems: Towards a theory of scientific growth*. Berkeley: University of California Press.

Levins, R. (1970). Extinction. In M. Gesternhaber (ed.), *Some Mathematical Problems in Biology* (pp. 77–107). Providence, RI: American Mathematical Society.

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the Entomological Society of America, 15,* 237–240.

Matt, G. E. & Cook, T. D. (2009). Threats to the validity of research syntheses. In H. M. Cooper & L. V. Hedges (*Eds.*) (2nd edition), *Handbook of research synthesis*. New York: Russell Sage.

Mill, J. S. (1843). *A system of logic, ratiocinative, and inductive*. London: Harrison and Co.

Mortensen, C. R. & Cialdini, R. B. (2009). Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass, 4*, 53-63.

Osburn, H. G. & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology, 77*, 115-22.

Platt, J. R, (1964). Strong inference. *Science, 146*, 347-353.

Rosenthal, R. & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Scott, A. G., Figueredo, A. J., and Hendrix, J. (1992). Meta-analysis of research in higher education retention studies. Paper presented at the American Evaluation Association Annual Meeting, Seattle.

Shadish, W. R. (1986). Planned critical multiplism: Some elaborations. *Behavioral Assessment, 8*, 75-103.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation, 60*, 13-57.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper and L. V. Hedges (*Eds.*), *The handbook of research synthesis*. New York: Russell Sage Foundation.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.