META-ANALYSIS: A PRIMER FOR LEGAL SCHOLARS

Jeremy A. Blumenthal[*]

Empirical research relevant to legal issues is common in other disciplines and is once again growing more common in the legal academy. Such research, however, varies widely in theoretical and methodological rigor and at times yields widely different results. Such disparate findings may bring into question the usefulness of such empirical research and may render it suspect in the eyes of practitioners, courts, and policy makers. One approach to helping address such concerns as well as other issues is meta-analysis*—the* quantitative*, rather than simply narrative review of empirical research. Meta-analysis synthesizes the relevant empirical literature, statistically summarizing the results of all empirical work in a particular area; and also identifies* moderator variables*, aspects of the various studies that might have influenced their findings. In this Article, I explain the importance of the meta-analytic approach, discussing what it is, why it is useful to members of the legal system, and the straightforward way of conducting a meta-analysis. The Article should be useful to legal academics, policy makers, courts, and practitioners.*

INTRODUCTION

Empirical legal scholarship is once again on the rise.[1] This return to empirical work, however, raises a number of concerns.[2] Legal scholars may be unfamiliar with the substance of another discipline, such as psychology,

---

* Syracuse University College of Law. J.D., University of Pennsylvania Law School; A.B., A.M., Ph.D., Harvard University. Special thanks to David C. Howell, Emeritus Professor, University of Vermont, for permission to use tables from his book. DAVID C. HOWELL, STATISTICAL METHODS FOR PSYCHOLOGY (3d ed. 1992). Appendices X and Z in the present Article were retyped with permission from portions of Appendices χ^2 and Z in that book. *Id.* at 637, 662-65. Values in Appendices RZR and ZRR in the present Article were computed by the author. I thank the Barclay Library Staff for substantial research assistance. I dedicate this Article to Matthew Stephen Bernstein Blumenthal.

1. An empirical approach to legal scholarship has been advocated at times throughout the twentieth century, with varying success. *See* Gregory Mitchell, *Empirical Legal Scholarship as Scientific Dialogue*, 83 N.C. L. REV. 167, 168-69 (2005) (noting legal scholars' calls for more empirical legal research). *See generally* JOHN HENRY SCHLEGEL, AMERICAN LEGAL REALISM AND EMPIRICAL SOCIAL SCIENCE (1995) (providing detailed review of early efforts to develop empirical legal work); Jeremy A. Blumenthal, *Law and Social Science in the Twenty-First Century*, 12 S. CAL. INTERDISC. L.J. 1, 7-22 (2002) (reviewing historical efforts by legal scholars and social scientists to empirically examine aspects of the legal system). Empirical legal scholars recently instituted a journal and a "blog" devoted to the publication and the discussion of Empirical Legal Studies. *See generally* J. EMPIRICAL LEGAL STUD., *available at* http://www.blackwellpublishing.com/ journal.asp?ref=1740-1453&site=1 (last visited Apr. 15, 2007) (journal); Empirical Legal Studies, http://www.elsblog.org/about.html (last visited Apr. 15, 2007) (blog).

2. *See generally* Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1 (2002) (outlining numerous concerns with existing body of empirical legal literature).

economics, criminology, sociology, or political science, that they wish to integrate into a prescriptive legal discussion. Such scholars may be unfamiliar with appropriate methodological design for empirical studies, means of data collection, or statistical analysis. Multiple empirical studies in a literature may lead to confusion for both scholars and policy makers about how to reconcile different findings or about what a literature may actually say, leading to quite different interpretations of, and inferences from, studies on the same topic. Of course, this can also lead to deliberate picking and choosing from multiple available studies, in order to proffer a particular point. Overreliance on (and misunderstanding of) statistical significance at the "magic" .05 level can obscure findings of importance, leading to publication bias, misrepresentation of a body of work, and flawed understandings of the implications of research. And failure to consider outside factors that might influence a study's findings may lead to a focus on misleading elements in explaining results, making inferences, and setting policy.

As the recent increase in quality empirical legal work demonstrates, however, none of these problems is insurmountable. In particular, one underutilized approach in particular that helps address a number of these problems is *meta-analysis*, a means of quantitatively synthesizing a body of empirical studies in order not only to summarize the whole of the research—to look at the forest, rather than individual trees—but also to identify "moderator variables," aspects of the various studies (date of publication, sample size, variables studied, analyses used, author's affiliation, specific research question examined) that might have reliably affected their outcomes. Taking this approach helps summarize an entire body of research, giving practitioners, academics, researchers, and policy makers the best view of the state of a literature; helps explain why that literature might look the way it does; and helps develop theories for further research in that area.[3]

Although meta-analysis is quite common in other areas of empirical research—the social and behavioral sciences, education, medical, and epidemiological research—it is rarely used or even considered in empirical legal research.[4] In part this is due to not recognizing the benefits of conducting a meta-

3. *See* Mitchell, *supra* note 1, at 187-88 (identifying value of meta-analysis for legal scholarship). As meta-analyst experts John Hunter and Frank Schmidt have pointed out:

> For decades policymakers seeking factual foundations for policy have looked to psychological and social science research. Until recently, they have been disappointed to find research literatures that were conflicting and contradictory. As the number of studies on each particular question became larger and larger, this situation became increasingly frustrating and intolerable. These problems stemmed from reliance on defective procedures for achieving cumulative knowledge: the statistical significance test in individual primary studies in combination with the narrative subjective review of research literatures. Meta-analysis principles have now correctly diagnosed this problem and, more important, have provided the solution.

John E. Hunter & Frank L. Schmidt, *Cumulative Research Knowledge and Social Policy Formulation: The Critical Role of Meta-Analysis*, 2 PSYCHOL. PUB. POL'Y & L. 324, 342-43 (1996).

4. Dan Orr & Chris Guthrie, *Anchoring, Information, Expertise, and Negotiation: New Insights from Meta-Analysis*, 21 OHIO ST. J. ON DISP. RESOL. 597, 612 (2006) ("Legal scholars routinely cite meta-analyses to support empirical claims they want to make about the legal world, but few law

analysis, in part to simply not knowing how to conduct one, in part to misconceptions in the courtroom and the legislature that lead to decreased receptivity to such synthesis, and in part due to unfamiliarity with a substantive literature in other fields (and thus a lack of recognition that research exists that could be synthesized and applied). In this Article, I present an overview of the procedure, explaining the *what*, *why*, and *how* of meta-analysis: what it is and how it addresses some of the problems identified above, why it is so useful for both practitioners and legal academics (of empirical and nonempirical stripes), and the surprisingly simple steps involved in how to conduct one.[5]

I. META-ANALYSIS: DESCRIPTION ("WHAT?")

In traditional empirical research, an investigator examines the association, either causal or correlational, between different variables. This may be through experimental research, in which specific factors are manipulated and participants are randomly assigned to the different manipulation conditions, or it may be through observational research, in which no random assignment or manipulation is involved, but associations between various variables or sets of variables are of interest.

In such research, the conventional approach is to collect data on individual units of measurement—individual people, cases, agencies, judges, jurors, juries, statutes, etc.—in order to address specific research questions. Do men and women differently consider certain social interactions to be sexual harassment?[6] Do juries with more than *X* white jurors give more death sentences than those with fewer?[7] Can jurors comprehend one set of death penalty instructions better than another, with implications for guided discretion in capital sentencing?[8] Do

---

reviews have actually published *original* meta-analyses." (footnote omitted)).

5. In part, then, the Article helps address one of Lee Epstein and Gary King's concerns, the lack of articles in the legal literature "devoted exclusively to solving methodological problems unique to legal scholarship." Epstein & King, *supra* note 2, at 6 n.19. The methodological "problems" that meta-analysis addresses are hardly unique to empirical legal scholarship, but they are common enough in that scholarship that the procedure will be quite useful.

6. *See generally, e.g.*, Louise F. Fitzgerald & Alayne J. Ormerod, *Perceptions of Sexual Harassment: The Influence of Gender and Academic Context*, 15 PSYCHOL. WOMEN Q. 281 (1991) (concluding that combination of severity or explicitness of incident and gender of perceiver had a bearing on whether perceiver would consider incident harassment); Tricia S. Jones & Martin S. Remland, *Sources of Variability in Perceptions of and Responses to Sexual Harassment*, 27 SEX ROLES 121 (1992) (concluding that notion of harassing behavior varies based on nature of behavior, gender of observer, and gender of target of behavior); Natalie J. Maloviich & Jayne E. Stake, *Sexual Harassment on Campus: Individual Differences in Attitudes and Beliefs*, 14 PSYCHOL. WOMEN Q. 63 (1990) (concluding that self-esteem, sex-role attitudes, and gender all had bearing on perception of harassment).

7. *See* William J. Bowers et al., *Death Sentencing in Black and White: An Empirical Analysis of the Role of Jurors' Race and Jury Racial Composition*, 3 U. PA. J. CONST. L. 171 (2001) (examining whether jury's racial composition influences sentencing phase of death penalty cases and concluding that chance of defendant receiving death sentence is increased when jury is predominantly white and defendant is black).

8. *See generally, e.g.*, Craig Haney et al., *Deciding to Take a Life: Capital Juries, Sentencing Instructions, and the Jurisprudence of Death*, 50 J. SOC. ISSUES 149 (1994) (concluding that jury death

white jurors punish white defendants more leniently than black defendants, or more leniently than do black jurors?[9] Do judges appointed by a Republican President rule differently on certain issues than do ones appointed by a Democrat?[10] Data are collected on participant behavior, and inferences are made either about the impact of the experimental manipulation (one set of instructions leads to fewer death sentences), or about the connection between the observed data and the independent variable originally selected (political affiliation of a judge's appointer is associated with rulings on Issues *X* and *Y* but not *Z*).

Such conventional or primary studies, examining individual units of analysis, can thus yield information about the presence of a particular phenomenon, its estimated strength, and relevant factors that might affect that presence or strength. The *meta-analytic* approach is similar, but is conducted at a different level of analysis: the individual units that *meta-analysis* involves are the empirical studies in a particular body of research.[11] Thus, meta-analysis quantitatively synthesizes and combines an empirical literature to examine associations between characteristics of each individual study and that study's outcomes.[12] More generally, a meta-analysis has at least three objectives: (1) to

---

penalty instructions not only failed to guide juries' decisions but distorted decision-making process); James Luginbuhl, *Comprehension of Judges' Instructions in the Penalty Phase of a Capital Trial: Focus on Mitigating Circumstances*, 16 LAW & HUM. BEHAV. 203 (1992) (concluding that old version of jury death penalty instructions in North Carolina resulted in confusion among jurors as to significance of mitigating circumstances).

9. Three reviews of this literature provide examples of the many studies examining this question: Ronald Mazzella & Alan Feingold, *The Effects of Physical Attractiveness, Race, Socioeconomic Status, and Gender of Defendants and Victims on Judgments of Mock Jurors: A Meta-Analysis*, 24 J. APPLIED SOC. PSYCHOL. 1315 (1994); Tara L. Mitchell et al., *Racial Bias in Mock Juror Decision-Making: A Meta-Analytic Review of Defendant Treatment*, 29 LAW & HUM. BEHAV. 621 (2005); Laura T. Sweeney & Craig Haney, *The Influence of Race on Sentencing: A Meta-Analytic Review of Experimental Studies*, 10 BEHAV. SCI. & L. 179 (1992).

10. *See generally, e.g.*, Max M. Schanzenbach & Emerson H. Tiller, *Strategic Judging Under the United States Sentencing Guidelines: Positive Political Theory and Evidence*, 23 J.L. ECON. & ORG. 24 (2007) (concluding that political orientation impacted criminal sentencing in cases where judges departed from federal sentencing guidelines); Cass R. Sunstein et al., *Ideological Voting on Federal Courts of Appeals: A Preliminary Investigation*, 90 VA. L. REV. 301 (2004) (concluding that judicial votes are influenced by individual judge's political affiliation as well as by judicial panel's political composition).

11. *See* David B. Wilson, *Meta-Analytic Methods for Criminology*, 578 ANNALS AM. ACAD. POL. & SOC. SCI. 71, 72 (2001) ("averaging across studies is analogous to averaging across individuals within a single study").

12. Hines v. Consol. Rail Corp., 926 F.2d 262, 273 n.7 (3d Cir. 1991) ("Meta-analysis involves pooling data from a number of different epidemiological studies (in order to enhance sample size) and comparing the results of those pooled data with the results produced by each study individually."); Lewis v. City of Chi., No. 98 C 5596, 2005 WL 693618, at *13 n.8 (N.D. Ill. Mar. 22, 2005) ("A meta-analysis is a statistical analysis of the results of a collection of individual studies to integrate and summarize their results."); Black v. Rhone-Poulenc, Inc., 19 F. Supp. 2d 592, 604 (S.D.W. Va. 1998) ("In short, a meta-analysis simply pools all of the data from many studies and treats them as one mega-study."); United States v. Nguyen, 793 F. Supp. 497, 512 n.23 (D.N.J. 1992) (quoting expert witness Professor Steven Penrod defining meta-analysis as "accepted method of analysis" that "combin[es] the results of independent studies in order to arrive at a general conclusion"); Gene V. Glass, *Primary,*

identify the presence or absence of an effect in an existing empirical literature; (2) to evaluate the strength of that effect, for instance by summarizing the average effect across a set or subset of studies; and (3) to identify *moderator variables*, elements of the various studies that might have reliably affected their outcomes. That is, first, a synthesis of an entire literature can give a better sense than single studies or simple anecdotes and assumptions of whether a relationship or effect is present.[13] This is especially the case when the literature contains studies with apparently discrepant results.[14] Second, though, if a relationship is present, quantitative meta-analytic review can easily determine the average effect size, leading to stronger grounds on which to make policy inferences. Third, when multiple studies do give different results—potentially leading to ambiguity about whether there is "truly" an effect, and perhaps to different emphasis by policy makers—analysis of moderator variables can help parse what factors lead to particular results rather than others.[15] As discussed further below, many of the advantages of meta-analysis stem from the fact that it takes a quantitative, rather than qualitative, approach to cumulating and synthesizing research.

II. META-ANALYSIS: JUSTIFICATION ("WHY?")

In this Part, I explain the background of meta-analysis in more detail, emphasizing its benefits and responding to some typical criticisms. In particular, I discuss how the procedure addresses some of the concerns identified above that arise in the context of empirical legal scholarship. I also identify a number of additional advantages meta-analysis has for legal scholars and practitioners, especially over traditional narrative reviews. For instance, a strong meta-analysis will collect published and unpublished studies from all the relevant disciplines, helping familiarize legal researchers and practitioners with new bodies of work,

---

Secondary and Meta-Analysis of Research, 5 EDUC. RES. 3, 3 (1976) (stating meta-analysis is "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings"); Raymond J.G.M. Florax et al., Meta-analysis: A Tool for Upgrading Inputs of Macroeconomic Policy Models 1 (Apr. 1, 2002) (unpublished manuscript, *available at* http://www.tinbergen.nl/discussionpapers/02041.pdf) (stating that meta-analysis is used to synthesize and summarize the results previously reported in literature on area of research). Of the cases' descriptions, the definitions in *Nguyen* and *Lewis* are, in fact, the more accurate—meta-analysis synthesizes the *results* of studies, not their raw *data*. Statistical problems may arise when the raw data of studies are pooled, rather than their results. *See, e.g.*, ROBERT ROSENTHAL, META-ANALYTIC PROCEDURES FOR SOCIAL RESEARCH 99-101 (rev. ed. 1991) (explaining that pooling data from various studies can yield results paradoxically outside the range of results of any individual study).

13. Blumenthal, *supra* note 1, at 40 ("This aggregation and averaging can lead to a more robust finding from which to infer policy than either an abbreviated selection of research findings or pure assumption or anecdote." (footnote omitted)).

14. *See* Frank L. Schmidt, *Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers*, 1 PSYCHOL. METHODS 115, 123 (1996) ("Applications of meta-analysis to accumulated research literatures have generally shown that research findings are not nearly as conflicting as we had thought and that useful general conclusions can be drawn from past research.").

15. See *id.* at 122-23 for an explanation of the benefits of meta-analysis when confronted with seemingly conflicting findings based on moderator variables.

new theoretical approaches, and new methodologies. Summary information about a body of work helps focus knowledge about that research. For a number of reasons quantitative rather than qualitative review enriches scholars' understanding of a discipline. Meta-analysis de-emphasizes reliance on statistical significance testing, which can easily mislead researchers into seeing an effect where it is absent, or ignoring one that is present. A thorough meta-analysis identifies and statistically tests moderator variables in order to identify factors that influence studies' results. And important theoretical and policy implications stem from the results of a meta-analysis that might be obscured in traditional narrative review. Each of these benefits is developed further below.

A. Importance of Synthesizing Research

Troves of data in other disciplines, as well as the burgeoning empirical literatures in disparate areas of legal academia, make summaries of existing research eminently useful beginning points, whether for academics, agencies, legislatures, judges, or practitioners.[16] If nothing else, they help distill or "translate" a body of research with which a secondary investigator might be unfamiliar.[17] Moreover, such integrative research reviews give broad and representative overviews of the existing research, but also serve to evaluate existing hypotheses, policy positions, or other assumptions. In particular,

> [w]hen the collective evidence is sufficiently abundant and clear in its conclusions, literature reviews can serve a falsifying function, demonstrating the incorrectness of one or more theoretical positions. When the evidence is less abundant or clear, literature reviews can serve the equally important function of clarifying the issues in the debate and directing research toward important open empirical issues that must be resolved to advance the debate.[18]

Moreover, by focusing on a broad range of studies, literature reviews can also help elucidate aspects of the studies that might have influenced different findings.[19] As discussed in more detail below, by statistically evaluating

16. *See* Harris Cooper & Larry V. Hedges, *Research Synthesis as a Scientific Enterprise*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS 3, 4-5 (Harris Cooper & Larry V. Hedges eds., 1994) (citations omitted) (noting common use of literature reviews as means of keeping up with developing empirical research); Richard Lempert, *"Between Cup and Lip": Social Science Influences on Law and Policy*, 10 LAW & POL'Y 167, 175 (1988) (suggesting that sources synthesizing studies are more useful to policy makers than single studies because they provide better overview of that area of research); Sarah H. Ramsey & Robert F. Kelly, *Social Science Knowledge in Family Law Cases: Judicial Gate-Keeping in the* Daubert *Era*, 59 U. MIAMI L. REV. 1, 79 (2004) (suggesting that reviews of literature on area of study will allow judges to be better informed when confronting and assessing specific findings of study presented in court). Unsurprisingly, literature reviews have thus long been among the most commonly cited articles, at least in the social sciences. Cooper & Hedges, *supra*, at 4.

17. Lempert, *supra* note 16, at 175; Wilson, *supra* note 11, at 85.

18. Gregory Mitchell, *Beyond Fireside Inductions*, 32 FLA. ST. U. L. REV. 315, 318 (2005).

19. *See, e.g.*, Lempert, *supra* note 16, at 176 (explaining that comparison of various studies, conducted differently, and sometimes reaching seemingly conflicting results, can produce explanations for inconsistencies and provide broader understanding of an area of research); Thomas O. McGarity, *Proposal for Linking Culpability and Causation to Ensure Corporate Accountability for Toxic Risks*, 26 WM. & MARY ENVTL. L. & POL'Y REV. 1, 65 (2001) (citation omitted) (noting that meta-analysis can

moderator variables, meta-analysis provides specific, focused ways of doing so.[20]

B. *Improvements Over Traditional Narrative Reviews*

Summaries of existing research ground a researcher or policy maker in a particular body of work. But a number of factors illustrate the advantage of meta-analyses over traditional narrative reviews. The comprehensive nature of meta-analysis—the consideration of an entire corpus of empirical studies in a research area—is one aspect that sets it apart from traditional synthesis. As a body of research grows larger, it is correspondingly difficult, of course, to include a narrative discussion of each relevant study.[21] Further, for many reasons, the narrative reviewer must make choices about which studies to mention at all, which studies to discuss at length, and how to characterize each study's different aspects and findings. If the grounds for such judgment calls are not made explicit, an uneven review may result.[22] Moreover, when exclusion criteria and definitions are not made explicit, subtle—or not so subtle—bias may creep into different reviews, with little opportunity for a reader to evaluate the reviewer's standards.[23] Gene Glass and colleagues, for instance, identified three narrative reviews of essentially the same set of studies, one finding "striking" effects of a treatment, one finding "little difference" in the same treatment, and one reporting no "firm conclusions" about it.[24] Meta-analysis—by definition—seeks to include, in quantitative form, *every* relevant study in a discipline, helping to avoid this concern over picking and choosing studies.

Even when a narrative reviewer *does* make those judgments explicit, however, excluding data because of perceived methodological or other deficiencies in fact fails to provide a full picture of the research. Meta-analysis, in contrast, tends to value almost any data as informative to some extent; as discussed below, the most common approach is to include allegedly "deficient" studies, but to quantitatively weigh them by their quality, sample size, or other factors. Indeed, in meta-analysis, "the influence of study quality on findings has

provide better understanding of a particular area of research as a whole than results from individual study).

20. See *infra* Part II.D for a discussion of the use of meta-analysis to examine moderator variables impacting individual studies.

21. *E.g.*, JOHN E. HUNTER ET AL., META-ANALYSIS: CUMULATING RESEARCH FINDINGS ACROSS STUDIES 26 (1982) (criticizing traditional review techniques for failing to integrate seemingly conflicting studies into review of area of research); John P.A. Ioannidis & Joseph Lau, *Systematic Review of Medical Evidence*, 12 J.L. & POL'Y 509, 534 (2004) (suggesting that quantitative methods of integrating data from many studies in particular research topic are necessary for more total appraisal of that topic).

22. *See* GENE V. GLASS ET AL., META-ANALYSIS IN SOCIAL RESEARCH 13 (1981) (listing various problems that arise when reviewers selectively review studies but fail to explain or assess the methods of their selections); FREDRIC M. WOLF, META-ANALYSIS: QUANTITATIVE METHODS FOR RESEARCH SYNTHESIS 10 (1986) (noting that literature reviews are often subject to biases of reviewers).

23. *See* Hunter & Schmidt, *supra* note 3, at 330 ("Relying on various personal and subjective theories and beliefs about methodological quality, reviewers often exclude[] all but a small number of studies as *methodologically inadequate* and then base[] their reviews on only the remaining few studies." (internal quotation marks omitted) (emphasis added)).

24. GLASS ET AL., *supra* note 22, at 18.

been regarded as an empirical a posteriori question, not an a priori matter of opinion or judgment used to exclude large numbers of studies from consideration."[25]

Further, narrative reviews typically do not evaluate in much detail the possible relationships between the studies' findings and various elements that might influence those findings—i.e., factors that might act as moderator variables.[26] Finally, practically speaking it is quite rare to find a free-standing law review summary of empirical research, whether qualitative or quantitative; when a review appears it is typically a short section of a longer article and thus even more subject to all of these space and selection concerns.

The distinction between narrative and quantitative reviews has itself been subjected to experimental research. Harris Cooper and Robert Rosenthal asked two groups of empirical researchers to evaluate seven individual empirical studies (testing a relationship between sex and the personality trait of persistence).[27] Both groups were presented with the original studies and asked to draw conclusions about the body of research. Participants in one group were asked to employ whatever evaluative criteria they would use if they were preparing a report for class or to submit as a manuscript. The second group was walked through meta-analytic procedures and was asked to report a quantitative summary of the findings. Despite cumulative evidence from all seven studies showing a relationship at conventional levels of statistical significance, three-quarters of those reviewers approaching the problem with traditional methods read the studies as showing no relationship.[28] Almost seventy percent of those using a quantitative approach, however, correctly identified the relationship.[29] In another instance, a meta-analysis of studies on a particular pedagogical method demonstrated a substantial effect, where a narrative review of the same literature a short time earlier had concluded that the method had no benefit.[30] Thus, taking a narrative approach to reviewing empirical scholarship risks losing relevant information; it also risks committing a Type II error—that is, mistakenly concluding that no effect exists when in fact it does. Clearly, policy inferences are vulnerable to such methodologically based errors.

---

25. *Id.* at 22.

26. *Id.* at 13.

27. Harris M. Cooper & Robert Rosenthal, *Statistical Versus Traditional Procedures for Summarizing Research Findings*, 87 PSYCHOL. BULL. 442 (1980).

28. *Id.* at 448.

29. *Id.* The two groups' estimates of the size and strength of the relationship differed as well. *Id.*

30. *Compare* Doris L. Redfield & Elaine W. Rousseau, *A Meta-analysis of Experimental Research on Teacher Questioning Behavior*, 51 REV. EDUC. RES. 237 (1981) (reporting meta-analytic study finding significant and "positive effect on student achievement" when higher cognitive questioning was used), *with* Philip H. Winne, *Experiments Relating Teachers' Use of Higher Cognitive Questions to Student Achievement*, 49 REV. EDUC. RES. 13 (1979) (reporting narrative review finding little beneficial pedagogical effect of teachers' asking students "higher cognitive" questions as opposed to "fact" questions).

C. *De-Emphasizing Statistical Significance*

Meta-analysis avoids a number of other common concerns as well. Perhaps most valuably, it de-emphasizes traditional statistical significance testing and the Holy Grail of a .05 *p*-value, in favor of measures of the strength of particular effects ("effect sizes").[31] Significance testing (also known as "null hypothesis statistical testing," or "NHST"[32]), perhaps the most widely known approach to evaluating research findings, can be misleading for a number of reasons.

First, significance testing does not tell us what we often think it does. A significance test yields a "*p*-value," a value between 0 and 1.00 that describes the likelihood the observed results would have occurred by chance, if there were no true difference between the experimental conditions.[33] Traditionally, a *p*-value of .05 (i.e., a five percent or 1 in 20 likelihood) is deemed "statistically significant" and thus worthy of attention.[34] Nevertheless, this is *not* the same as suggesting that applying a .05 level of statistical significance gives correct results ninety-five percent of the time,[35] a common (mis)interpretation. Second, a significance level is entirely dependent on the size of the sample studied.[36] Two studies examining precisely the same research question and using precisely the same methodology, but using different-sized samples, might arrive at discrepant conclusions because the significance levels they elicit will differ. Basing policy on one or the other of those studies might thus be misleading or misguided. Third,

> [f]ailure to reach this 'magical' .05 level . . . does not mean that a difference is not meaningful: imagine, for instance, a weather report that there is a 95% chance of rain. It would be surprising if a change in that report to only a 94% chance (*i.e.*, a p-value of .06),[37] or even a 90% chance (*i.e.*, a p-value of .10), of rain would tend to lead people to leave their umbrellas home.[38]

---

31. R. Rosenthal & M.R. DiMatteo, *Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews*, 52 ANN. REV. PSYCHOL. 59, 63 (2001) ("Meta-analysis prevents our reliance on the significance test of any one finding as a measure of its value"); Robert Rosenthal, *Writing Meta-Analytic Reviews*, 118 PSYCHOL. BULL. 183, 185 (1995) ("Effect size estimates are the meta-analytic coin of the realm.").

32. In a well-known critical discussion of such testing, Jacob Cohen elaborated on the acronym saying, "I resisted the temptation to call it statistical hypothesis inference testing." Jacob Cohen, *The Earth is Round (p < .05)*, 49 AM. PSYCHOL. 997, 997 (1994).

33. Jeremy A. Blumenthal, *Does Mood Influence Moral Judgment? An Empirical Test with Legal and Policy Implications*, 29 LAW & PSYCHOL. REV. 1, 9 n.54 (2005).

34. *Id.*; *see also* DAVID C. HOWELL, STATISTICAL METHODS FOR PSYCHOLOGY 88 (3d ed. 1992) (explaining significance level test, which discards null hypotheses with probabilities less than or equal to .05).

35. HUNTER ET AL., *supra* note 21, at 20.

36. *E.g.*, ROSENTHAL, *supra* note 12, at 14 (explaining that test of significance is product of effect size—or strength of relationship between variables—and study size); Rosenthal & DiMatteo, *supra* note 31, at 63 (stating that significance of effect size is determined by size of study).

37. The analogy is inexact, but the broader point should be clear.

38. Blumenthal, *supra* note 33, at 9 n.54 (emphasizing importance of not solely relying on .05 *p*-values); *see also* Vuyanich v. Republic Nat'l Bank of Dallas, 505 F. Supp. 224, 272 (N.D. Tex. 1980) (noting problems with arbitrarily excluding results with *p*-value greater than .05 when such studies

By focusing on effect sizes instead, meta-analysis helps avoid such pitfalls. But even when significance testing is used, however, meta-analysis illustrates the importance of cumulating research findings rather than looking at them in isolation. At least one court seems to have excluded a meta-analysis from evidence simply because the findings of the underlying primary studies did not reach statistical significance.[39] This is simply inappropriate. Meta-analysis demonstrates that repeated results in the same direction, even if not significant at traditional levels, cumulate to more persuasive evidence of an effect than even a single, highly significant finding. For instance, as a mathematical matter, two studies that "only" reach significance at the $p = .06$ level are, cumulated, stronger evidence of an effect than a single study showing $p = .05$. Similarly, ten studies showing significance at "only" $p = .10$ are stronger evidence that the null hypothesis of no difference is false, than five studies showing $p = .05$.[40] "Meta-analysis thus provides the opportunity for even small and nonsignificant effects to contribute to the overall picture of the results of a research enterprise."[41]

Commentators have warned that the "lack of careful and regular attention by social scientists and lawyers to research findings that do not yield results significant at the .05 level makes for bad science, bad ethics, and uninformed uses of social science in the courtroom."[42] Emphasis on effect sizes rather than *p*-values can help avert the tendency for courts and commentators to reject findings that do not reach the magic .05 level,[43] and meta-analysis is an important

---

may still be useful), *vacated on other grounds*, 723 F.2d 1195 (5th Cir. 1984); Robert Rosenthal & Donald B. Rubin, *Comparing Significance Levels of Independent Studies*, 86 PSYCHOL. BULL. 1165 (1979) (noting that in reality, there is little difference between *p*-value of .05 and .06); *cf.* Robert Rosenthal & John Gaito, *The Interpretation of Levels of Significance by Psychological Researchers*, 55 J. PSYCHOL. 33 (1963) (empirically documenting overreliance among empirical scientists on .05 *p*-levels); Robert Rosenthal & John Gaito, *Further Evidence for the Cliff Effect in the Interpretation of Levels of Significance*, 15 PSYCHOL. REP. 570 (1964) (documenting same).

39. *See* Allison v. McGhan Med. Corp., 184 F.3d 1300, 1315 (11th Cir. 1999) (upholding district court's exclusion of expert's meta-analysis where the court found the "study unreliable because it was a re-analysis of other studies that had found no statistical correlation"). *But see In re* Paoli R.R. Yard PCB Litig., 916 F.2d 829, 856-58 (3d Cir. 1990) (reversing district court admissibility ruling that synthesizing nonsignificant studies could not achieve statistical significance).

40. Rosenthal & DiMatteo, *supra* note 31, at 63.

41. *Id.*; *see also* GLASS ET AL., *supra* note 22, at 220-21 (explaining that many studies with different weaknesses, when pulled together, can yield strong results).

42. Robert Rosenthal & Peter David Blanck, *Science and Ethics in Conducting, Analyzing, and Reporting Social Science Research: Implications for Social Scientists, Judges, and Lawyers*, 68 IND. L.J. 1209, 1220 (1993).

43. *See, e.g.*, DeLuca v. Merrell Dow Pharm., Inc., 911 F.2d 941, 947 (3d Cir. 1990) (citation omitted) (stating that "if P is greater than 5% the relationship is rejected as insignificant"); Coates v. Johnson & Johnson, 756 F.2d 524, 537 n.13 (7th Cir. 1985) ("A P value below .05 is generally considered to be statistically significant."); Presseisen v. Swarthmore Coll., 442 F. Supp. 593, 617 (E.D. Pa. 1977) ("Since a small value of P, i.e., less than .05 (for example, .04, .03, etc.), indicates an effect is statistically significant; and since all the P values in [question] are not less than .05, the average differences mentioned above could be attributable to chance alone."); Bloomquist v. Wapello City, 500 N.W.2d 1, 5 (Iowa 1993) (quoting *Deluca*, 911 F.2d at 947) (suggesting that relationship is regarded as insignificant for *p*-values greater than .05); Celestial S.D. Cassman & Lisa R. Pruitt, *A Kinder, Gentler Law School? Race, Ethnicity, Gender, and Legal Education at King Hall*, 38 U.C. DAVIS L. REV. 1209, 1242 & n.142 (dividing study results between those with statistically significant *p*-

means of maintaining this emphasis. Indeed, this approach, arguably, better comports with the legal system's perspective on proof.[44]

D. *Moderator Variables*

One of the most important aspects of the meta-analytic approach is the ability to systematically compare and contrast across studies, not simply average their effect sizes.[45] That is, a variety of methodological and substantive aspects of the various studies in a discipline might be associated with the effect sizes they report.[46]

Thus, again analogizing to single studies, various characteristics of the individual subjects (e.g., individual jurors) are often of interest in how they might affect the outcome variables. A simple example is whether a mock juror's race (or that of a defendant) might influence her decision making about the defendant.[47] But when the body of this race literature is analyzed in toto, factors relating to each study might appear that explain some of the variation in findings.[48] For instance, although a meta-analysis of such studies lent overall support to the belief that race influenced sentencing decisions, specific factors were identified that clarified why some studies yielded stronger effects than others.[49] The authors found that if the study was conducted outside the Southern United States, for instance, there was a tendency for the effect of the defendant's race to be larger. Similarly, studies that specified the race of the subject or the victim yielded larger effect sizes.[50]

As another example, a meta-analysis examining sex differences in perceptions of sexual harassment found that when a situation was presented to subjects by videotape (presumably more realistically), sex differences were far stronger than when subjects were simply asked via a phone or mail survey whether a certain behavior constituted sexual harassment.[51] It also found that

---

values of less than .05 and those with no detectable difference, or *p*-values of more than .05); Fred O. Smith, Jr., Note, *Gendered Justice: Do Male and Female Judges Rule Differently on Questions of Gay Rights?*, 57 STAN. L. REV. 2087, 2113-14 (2005) (noting, without additional information, that *p*-value was "out of the bounds of statistical significance at a .05 level" and that there was "not necessarily" a difference in groups being examined).

44. Richard D. Friedman, *The Death and Transfiguration of* Frye, 34 JURIMETRICS J. 133, 148 (1994) ("As compared to traditional methods of significance testing, meta-analysis may more closely approach our evidentiary system's willingness to allow an inference to be drawn from various bits of information, none of which independently supports the inference.").

45. Blumenthal, *supra* note 1, at 42 (discussing different types of moderator variables that could be considered under meta-analytic approach).

46. Mark W. Lipsey, *Those Confounded Moderators in Meta-Analysis: Good, Bad, and Ugly*, 587 ANNALS AM. ACAD. POL. & SOC. SCI. 69, 69 (2003).

47. *See generally* Bowers et al., *supra* note 7, at 181-89 (collecting examples of such research).

48. *E.g.*, Sweeney & Haney, *supra* note 9, at 179 (noting that methodological differences across numerous studies accounted for stronger results in some studies than others).

49. *Id.*

50. *Id.* at 190.

51. Jeremy A. Blumenthal, *The Reasonable Woman Standard: A Meta-Analytic Review of Gender Differences in Perceptions of Sexual Harassment*, 22 LAW & HUM. BEHAV. 33, 44 tbl.7 (1998).

"recent studies appear to report *larger* gender differences than earlier studies, not smaller as one might expect from increased awareness of the problem of sexual harassment."[52] A court's, legislator's, or commentator's choice to rely more heavily on either more or less recent empirical work in that area, depending on familiarity or accessibility, might thus have very real policy implications.

E. Policy Implications

Some commentators have argued that because of these benefits, presenting information in the form of meta-analyses is in fact the best way to serve courts and policy makers.[53] Over and above those advantages, though, there are potential direct benefits for policymaking.

First, meta-analysis allows (or forces) consumers of the research in a discipline to work from a common set of studies and from a uniform starting point that summarizes and analyzes the available research. Avoiding cherry-picking among various studies, especially when the number of studies is large, not only emphasizes the importance of looking at the whole body of research, but it thus also encourages transparency, potentially calling into question a party's decision to look at individual studies rather than at the broader synthesis. This latter concern was at issue in two important United States Supreme Court cases: *Lockhart v. McCree*[54] and *General Electric Co. v. Joiner*.[55]

In *Lockhart*, a case concerning the effects of death qualification on jurors' conviction proneness, then-Chief Justice Rehnquist sharply criticized social psychological literature reviewed in an amicus brief submitted by the American Psychological Association ("APA").[56] Despite the APA's assertions that the studies involved were methodologically sound, the Chief Justice discussed and dismissed each of the studies on a variety of methodological grounds.[57] Were a

52. *Id.* at 46.

53. *E.g.*, Blumenthal, *supra* note 1, at 43-44 (asserting that benefits of meta-analysis comport with goals of judges and social scientists in yielding practical legal applications); Ioannidis & Lau, *supra* note 21, at 535 (noting that meta-analytical analysis can aid judges and lawyers in understanding expert testimony).

54. 476 U.S. 162 (1986).

55. 522 U.S. 136 (1997).

56. *Lockhart*, 476 U.S. at 169-73.

57. *Id.* Other courts have reacted similarly. *See* Free v. Peters, 12 F.3d 700, 705-06 (7th Cir. 1993) (criticizing studies on juror comprehension of sentencing instructions); State v. Deck, 994 S.W.2d 527, 542-43 (Mo. 1999) (en banc) (same), *aff'd in part and rev'd in part*, 68 S.W.3d 418 (Mo. 2002), *rev'd on other grounds*, 544 U.S. 622 (2005); *cf.* Green v. United States, 126 S. Ct. 497, 497 (2005) (denying certiorari in federal death penalty case in which circuit court reversed trial judge's order, based on empirical evidence of death-qualified individuals' bias, to bifurcate guilt and sentencing jury).

Then-Chief Justice Rehnquist's disparagement of the literature has since been criticized in turn by a number of scholars. *E.g.*, Phoebe C. Ellsworth, *Unpleasant Facts: The Supreme Court's Response to Empirical Research on Capital Punishment*, *in* CHALLENGING CAPITAL PUNISHMENT: LEGAL AND SOCIAL SCIENCE APPROACHES 177, 196-97 (Kenneth C. Haas & James A. Inciardi eds., 1988) (arguing that due to unanimity of results of various studies, Rehnquist's perceived flaws in methodology should not have been fatal flaws); J. Alexander Tanford, *The Limits of a Scientific Jurisprudence: The Supreme Court and Psychology*, 66 IND. L.J. 137, 145-47 (1990) (suggesting that Rehnquist, with no

meta-analysis presented synthesizing the entire body of knowledge, it may have been more difficult to distinguish the individual studies in that way. Indeed, a subsequent meta-analysis showed that "[s]umming across multiple studies using different trials, methods, and participants, the evidence remains firm: Death qualified jurors are more conviction prone than nondeath qualified jurors."[58]

Similarly, in *Joiner*, the Chief Justice reviewed a number of studies one by one to show that they were "so dissimilar to the facts presented in this litigation" that excluding them from trial was appropriate.[59] In contrast, Justice Stevens, in his partial dissent from *Joiner*, emphasized the usefulness of viewing "all the studies taken together," rather than parsing them one by one.[60] Meta-analytic presentation of that information might (conceivably) have led to different admissibility rulings.[61]

Another policy benefit of meta-analysis is the potential identification of small effects. Failure to reach statistical significance—even when an effect size is of substance—or, alternatively, effect sizes that are statistically significant but are small, can lead to inferences that the effect is unimportant.[62] Determining the practical, rather than the statistical, significance of an effect is of course essential—even large effects can be of little practical importance. But finding a small but robust effect across a large body of empirical work can lend credence to claims that policy makers should devote attention to the findings.[63]

Finally, an important policy implication of meta-analytic findings is to

---

training in social sciences, should have given more credence to findings of experimental psychologists); William C. Thompson, *Death Qualification After* Wainwright v. Witt *and* Lockhart v. McCree, 13 LAW & HUM. BEHAV. 185, 195-98 (1989) (criticizing Rehnquist's dismissal of studies he saw as less than definitive, even though they could have been nevertheless informative).

58. Joseph W. Filkins et al., *An Evaluation of the Biasing Effects of Death Qualification: A Meta-Analytic/Computer Simulation Approach*, *in* THEORY AND RESEARCH ON SMALL GROUPS 153, 171 (R. Scott Tindale et al. eds., 1998). Filkins and colleagues qualified their statement by noting that to some extent the *Lockhart* Court was correct; based on their computer simulations of jury decision making, death-qualified *juries* (i.e., rather than *jurors*) are less likely to convict than had been thought, though still slightly more so than non-death-qualified juries. *Id.* at 171-72.

59. *Joiner*, 522 U.S. at 144-45.

60. *Id.* at 153 (Stevens, J., concurring in part and dissenting in part).

61. *Cf.* Blumenthal, *supra* note 1, at 41 (suggesting that Stevens's approach to *Joiner* studies might have been able to yield more useful conclusions than Rehnquist's approach). Or it might not. The majority and Justice Stevens both dismissed the underlying studies because they did not reach conventional statistical significance levels. *See Joiner*, 522 U.S. at 154 & n.8 (Stevens, J., concurring in part and dissenting in part) (conceding that even broader meta-analysis approach yielded statistically insignificant results in this case). As noted above, however, this is misguided; a meta-analysis might have shown that overall, the findings across all studies did reach such levels.

62. *E.g.*, Stephen J. Deery & Roderick D. Iverson, *Labor-Management Cooperation: Antecedents and Impact on Organizational Performance*, 58 INDUS. & LAB. REL. REV. 588, 605 (2005) ("It is important, however, not to overstate the practical implications of our findings. We could explain only a relatively small proportion of the variance in the performance measures.").

63. *Cf.* Jerome McCristal Culp, Jr., *Toward a Black Legal Scholarship: Race and Original Understandings*, 1991 DUKE L.J. 39, 87 n.142 ("Statistically, a model with a very small R-squared may be a better model than one which purports to explain all of the variance."). For further discussion about presenting meta-analytic and other findings in a way that demonstrates the practical importance of even small effects, see *infra* Part III.D.

prompt further research and identify profitable contexts for such research. For instance, robust findings and methods can be identified, as can those that are more tentative, so that subsequent researchers can focus on one or the other.[64] Further, by identifying moderator variables, meta-analysis can prompt research that can refine and elaborate on theoretical, methodological, or pragmatic factors with different influences on the empirical findings.

F. Objections

The meta-analytic approach is hardly a panacea for every concern about empirical scholarship.[65] And, of course, it should not replace primary research studies—obviously, no meta-analysis would be possible without them. But in planning research and making policy inferences, the meta-analytic approach can be a substantial improvement over a simple focus on individual studies[66] and, especially, over the traditional narrative reviews of empirical research that appear in law journals.

Nevertheless, objections to meta-analysis arise. Few of the criticisms, however, are ultimately persuasive, and researchers should not be discouraged from pursuing such syntheses and presenting them to courts and policy makers. A number of criticisms are in fact applicable to traditional reviews. Some stem from misunderstanding the purpose and goals of the meta-analytic endeavor. Others, while relevant, are typically considered and addressed in undertaking a meta-analysis.

1. Including "Bad" Studies

One common objection is that the very inclusiveness of meta-analysis leads to consideration of studies of poor quality, with the suggestion that any results will therefore be tainted (the "garbage in, garbage out" hypothesis).[67] Although this is partly correct,[68] it is also so for narrative reviews, where it can be even less clear what studies are of better or worse quality.[69] It is also the case that having more data is better than having fewer[70]: for instance, increased power from

64. *E.g.*, Wilson, *supra* note 11, at 85 (noting that meta-analysis can be used to identify areas of study that are well researched and areas where more research is needed).

65. Blumenthal, *supra* note 1, at 45.

66. Lempert, *supra* note 16, at 176-77 (cautioning policy and legal decision makers against overreliance on single-study analyses).

67. *See* MORTON HUNT, HOW SCIENCE TAKES STOCK: THE STORY OF META-ANALYSIS 42 (1997) (criticizing meta-analysis for integrating studies of poor as well as sound design).

68. *See* Rosenthal & DiMatteo, *supra* note 31, at 66-67 (acknowledging "'garbage in and garbage out' issue" and suggesting weighting technique to accommodate quality variance among studies.

69. It may be too much to say that "bad" studies will be defined as those of our "enemies," see ROSENTHAL, *supra* note 12, at 130 (citation omitted), but subjectivity in such evaluation will be more common in narrative reviews.

70. Jessica Gurevitch & Larry V. Hedges, *Statistical Issues in Ecological Meta-Analyses*, 80 ECOLOGY 1142, 1146 (1999) (stating that developing methods for integrating poorly reported data is more desirable than ignoring those data altogether). The same is true for the synthesis of data of varying quality. *See* Sarah H. Ramsey & Robert F. Kelly, *Using Social Science Research in Family Law Analysis and Formation: Problems and Prospects*, 3 S. CAL. INTERDISC. L.J. 631, 680 (1994) (noting

including more studies helps narrow the estimate of the "true" effect size of the phenomenon in question. Moreover, it is possible to quantify study quality as simply another factor that might affect a study's results.[71] Finding that there is a relationship allows quantification of that influence and also gives justification for weighting higher-quality studies more heavily.[72] Of course, finding that there is *no* relationship between study quality and observed effect size is helpful as well.[73]

2. Heterogeneity (or Homogeneity) of Studies

A related objection is that studies included in a meta-analysis may be too dissimilar in methodology, variables examined, or other factors to compare. This is, again, a criticism equally applicable to narrative reviews. But it can also be more of a boon than a detriment, especially (perhaps counterintuitively) in terms of generalizability. That is, the objection runs that a meta-analyst is comparing "apples and oranges," and thus cannot generalize past either type. Indeed, some courts follow such logic to some extent.[74]

As others have pointed out, though, comparing apples and oranges is quite beneficial if what one is seeking to do is to generalize to fruit.[75] In other words, precise replication with exactly analogous participants and identical variables is typically not of primary interest in seeking to empirically identify a general effect. Thus, heterogeneity in methodology and other aspects of the studies may be beneficial because cumulation of "conceptual replications can show that a relationship is observed across a range of methodological and substantive variability."[76] Similarly, if one is willing to accept the generalization across subjects that occurs within individual studies, one should be as willing to accept that which takes place across studies, where the statistical power and accuracy can be that much greater.[77] Moreover, as with many of the factors describing the different studies, methodology or other differences can be incorporated

that many advocates believe meta-analysis can make good use of methodologically poor studies and maintaining that those studies should not always be excluded from reviews).

71. *E.g.*, WOLF, *supra* note 22, at 15 (explaining meta-analysis' ability to empirically handle quality of research design across various tests). See also *supra* note 25 and accompanying text for a discussion of assessing study quality and incorporating those assessments into a meta-analysis.

72. ROSENTHAL, *supra* note 12, at 130.

73. Indeed, there is evidence that in the typical meta-analysis, there is little such relationship. *See* GLASS ET AL., *supra* note 22, at 226 (stating that in reviews using large enough number of cases, differences in effects of high-quality and low-quality studies were small). Thus, dropping studies based on perceived methodological deficiencies wastes data. *Cf.* Nancy G. Berman & Robert A. Parker, *Meta-Analysis: Neither Quick Nor Easy*, 2 BMC MED. RES. METHODOLOGY 10, 17-18 (2002) ("[G]iven the effort that goes into identifying and evaluating papers, ignoring or rejecting valuable information is wasteful.").

74. *E.g.*, McNeil-P.C.C., Inc. v. Bristol-Myers Squibb Co., 938 F.2d 1544, 1547 (2d Cir. 1991) ("Moreover, evidence at trial showed that only seven out of the thirty studies incorporated into the meta-analysis concerned acetaminophen. [The district court judge] therefore accorded the meta-analysis little or no weight.").

75. Gene V. Glass, *In Defense of Generalization*, 1 BEHAV. & BRAIN SCI. 394, 395 (1978).

76. Wilson, *supra* note 11, at 73.

77. ROSENTHAL, *supra* note 12, at 129.

empirically as variables to be tested for a relationship with the observed effect size outcomes.[78]

An opposite sort of concern might also be broached—the objection might be that a *lack* of independence among studies biases the results. This might especially be the case in the context of econometric or multiple regression studies—that is, there may be a number of different studies constructing equations with different variables but sampling the same data set (e.g., all Supreme Court opinions, or multiple time series regression studies that use much the same data but update by a year or two).[79] Alternatively, there may be a number of related studies or experiments in a particular area, all conducted by the same researcher or team of researchers; this might lead to a concern that the approach or findings may not be independent.[80]

To address the latter concern, it is quite simple to incorporate author or research team as one of the moderator variables to consider statistically.[81] Addressing the former has been discussed in somewhat more detail, and may be more of a "gray area," especially in the time series case.[82] On balance, though, as one commentator has pointed out, even different studies that use the same or similar *data*, but different methodologies, equations, or model specifications, are likely sufficiently independent to include in the same meta-analysis: "A set of data does not contain one right answer, but rather a distribution of *plausible* estimates. This distribution is a function of largely random (mis)specification errors."[83] Repeating analysis of the same or similar data sets can in fact help narrow that distribution, improving the specificity of the overall analysis and estimates.[84] A meta-analyst, of course, should identify when multiple such studies are included in a review.[85]

3. Nonindependence of Results

Another sort of nonindependence might stem from experimental research, where respondents give responses on multiple dependent variables. It may also come from observational research, where a researcher, for instance, conducts a number of multiple regressions that differ by only one or a few predictors (more generally, this is relevant any time a study gives more than one effect size estimate). In the former case, it is easy enough to conduct multiple meta-

78. *E.g.*, WOLF, *supra* note 22, at 15 (asserting that even problematic differences in methodology can be examined empirically).

79. *See* Robert S. Goldfarb & H.O. Stekler, *Meta-Analysis*, 16 J. ECON. PERSP. 225, 225 (2002) (noting such concerns).

80. *E.g.*, ROSENTHAL, *supra* note 12, at 131 (noting that studies conducted by same laboratories or research groups may end up bearing similarities and may not be independent).

81. *See* Rosenthal & DiMatteo, *supra* note 31, at 67 ("It is possible and often valuable to block by laboratory or researcher and examine this as a moderator variable.").

82. T.D. Stanley, *Response from T.D. Stanley*, 16 J. ECON. PERSP. 227, 227 (2002).

83. *Id.* at 227-28.

84. *Id.* at 228.

85. *E.g.*, Blumenthal, *supra* note 51, at 43 (noting three studies that made use of interviews from the same subjects).

analyses on the separate dependent variables.[86] The latter case is more interesting, and there are at least two approaches to addressing it.

The first emphasizes that nonindependence is not a *statistical* flaw or bias, as it can be in significance testing or other statistical analysis—that is, having correlated results does not statistically bias a meta-analysis in the same way that it violates the relevant significance testing assumptions. Rather, including multiple effect sizes from the same study in a meta-analysis simply weights that study proportionately to the number of effect sizes that study generates.[87] The meta-analyst may choose not to do so, of course, and may choose the second approach, to instead average the multiple effect sizes from a particular study. Using the ordinary mean of the multiple effect size estimates is a conventional and somewhat conservative approach; less conservative approaches are also available.[88]

4. Publication Bias

Another concern—though again, one that is as relevant to narrative reviews—suggests that focusing on published articles biases the review, because of the tendency to only publish statistically significant results reaching the "magical" .05 *p*-level. This concern is valid; three approaches might be taken to ameliorate it. One is institutional and highly quixotic: journals should be open to publishing not only statistically significant findings, but null findings as well. Relatedly, and only somewhat less quixotic: journals should be open to findings that do not quite reach .05 levels, but that nevertheless obtain important and interesting findings. Researchers who emphasize effect sizes, rather than significance levels, will help with both of these points by drawing attention away from misleading *p*-levels.

The other two approaches are more practical and more directly relevant to the meta-analytic procedures. The first, as sketched further below in Part III.A, is to be sure that the meta-analyst's gathering of the existing work is not limited to published articles.[89] Not only symposium or conference proceedings, but also working papers and studies that were conducted but that did not find significant results, should be included in the review—the point of a meta-analysis is to obtain all relevant data. Publication status may in fact be used as a moderator variable, to examine whether publication bias does exist in the set of studies.

The last approach is statistical. Here, publication bias is presumed—that is,

86. *See, e.g.*, ROSENTHAL, *supra* note 12, at 26 (suggesting that for study of effects of alcoholism treatment programs, separate analyses could be performed for dependent variables like sobriety, number of days of employment or arrests, general medical health, and personal and social adjustment).

87. *Id.* at 27.

88. *See* Robert Rosenthal & Donald B. Rubin, *Meta-Analytic Procedures for Combining Studies with Multiple Effect Sizes*, 99 PSYCHOL. BULL. 400, 401-02 (1986) (discussing procedures for combining multiple effect sizes from single studies).

89. *E.g.*, WOLF, *supra* note 22, at 15 (advising that to combat bias in favor of significant results in published articles, researcher should review results in books, dissertations, and unpublished papers from professional meetings).

the reviewer assumes that published articles represent the five percent of studies where an effect was observed despite its actual absence (Type I errors) while the other ninety-five percent of studies showing nonsignificant results are wasting away in researchers' file cabinets.[90] This "file drawer problem" can be addressed statistically by calculating how many of those null result studies must be tucked away in file cabinets before casting doubt on the existing corpus of studies. That is, a meta-analyst can calculate how many such studies must exist, unpublished, in order to suggest that the observed findings occurred solely by chance.[91] This sort of discussion can be helpful in anticipating criticism of a meta-analytic review. For instance, in a meta-analysis examining the relationship between media violence and antisocial behavior, Haejung Paik and George Comstock calculated that for some of their effects, hundreds of thousands of studies finding null results would have to have been conducted, but not published or otherwise disseminated, in order to bring into question their findings of a relationship.[92]

5. Other Objections

The criticisms sketched above are among the more common,[93] but are also ones that are considered and accommodated in the meta-analytic tradition. Other objections, including ones by courts and commentators, are less troubling, as they typically display unfamiliarity or misunderstanding of the meta-analytic enterprise.[94]

For instance, a recent description of meta-analysis made the mistake that some courts did in describing it as a collation of *data* from a set of studies (rather than a synthesis of results) and suggested that meta-analysis is "highly susceptible to bias."[95] Although various judgments about the set of studies included are made in the review process, meta-analysis, more than traditional narrative review, is much more likely to articulate and even quantify the judgment criteria involved. Even more oddly, that review suggested that meta-analyses "are rarely published because reviewers tend to be especially harsh to

90. *See, e.g.*, Robert Rosenthal, *The "File Drawer Problem" and Tolerance for Null Results*, 86 PSYCHOL. BULL. 638, 638 (1979) (describing "file draw problem" in which majority of studies did not have significant—for example, $p > .05$—results and thus are not found within journals); Robert Rosenthal & Donald B. Rubin, *Comment: Assumptions and Procedures in the File Drawer Problem*, 3 STAT. SCI. 120, 120 (1988) (observing that assumption underlying original file draw computations is that selection process results in studies with significant results being published and retrieved, while nonsignificant results are not published or retrieved).

91. A straightforward guide to such calculations is available in ROSENTHAL, *supra* note 12, at 104-05.

92. Haejung Paik & George Comstock, *The Effects of Television Violence on Antisocial Behavior: A Meta-Analysis*, 21 COMM. RES. 516, 530 tbl.4 (1994).

93. Additional discussion of potential limitations of meta-analysis and other research synthesis may be found in Harris Cooper & Larry V. Hedges, *Potentials and Limitations of Research Synthesis*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS, *supra* note 16, at 521, 523-24, and in Georg E. Matt & Thomas D. Cook, *Threats to the Validity of Research Synthesis*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS, *supra* note 16, at 503, 503.

94. *See* Blumenthal, *supra* note 1, at 38-46 (identifying and discussing such concerns).

95. Patricia E. Lin, Note, *Opening the Gates to Scientific Evidence in Toxic Exposure Cases: Medical Monitoring and* Daubert, 17 REV. LITIG. 551, 581 (1998).

the methodology."[96] This is simply false—hundreds of meta-analyses are published annually in all areas of the social, medical, and other sciences.[97] That author's perspective reflects one that appears in some court cases as well—the suggestion that meta-analysis is a "novel" procedure.[98] It is not.[99]

III. META-ANALYSIS: MECHANICS ("HOW?")

The discussion above outlined *what* a meta-analysis is and *why* empirical legal scholars, courts, practitioners, and policy makers should find the approach of interest and importance. I now turn to some mechanics of meta-analysis—the *how*. I outline briefly how a researcher might collect the literature to be synthesized, discuss coding and evaluation of that literature, and then turn to the straightforward mathematical computations involved. This Part concludes with a brief discussion of how meta-analytic results might usefully be presented.

A. Literature Review and Selection

As an initial matter, of course, the body of empirical literature to be synthesized must be identified. This becomes more than simply a Westlaw or Lexis search of case law or law journal articles, as much relevant empirical work will be located in the social science context. Depending on the research question at issue, a reviewer will need to canvass journals in psychology, sociology, economics, political science, criminology, or other disciplines, as well as the relevant legal publications.[100] Electronic databases exist for such searches as well, though they are less often full-text searchable.[101]

The search should not be limited to published research, however. As suggested above, such a focus might inappropriately bias the review because published work tends to report results significant at conventional levels; those who do not find statistically significant results are either unable to publish or unwilling to try. Nevertheless, with the goal of not wasting existing data, a meta-

96. *Id.* at 581.

97. *See In re* Paoli R.R. Yard PCB Litig., 916 F.2d 829, 857 (3d Cir. 1990) ("[H]undreds of meta-analyses are done each year . . . ."); ROSENTHAL, *supra* note 12, at 10-11 (listing early examples). Meta-analyses being "rarely published" in law journals is more likely due to their being rarely conducted and submitted.

98. *E.g.*, *In re Paoli R.R. Yard PCB Litig.*, 706 F. Supp. at 373 (accepting defense expert's contention that meta-analysis is novel scientific technique that is to be evaluated under *United States v. Downing*, 753 F.2d 1224 (3d Cir. 1985)).

99. ROSENTHAL, *supra* note 12, at 5-7 (reviewing early examples of meta-analysis); Cooper & Hedges, *supra* note 16, at 5-6 (reviewing historical examples); Ingram Olkin, *History and Goals*, *in* THE FUTURE OF META-ANALYSIS 3, 3-9 (Kenneth W. Wachter & Miron L. Straf eds., 1990) (noting examples throughout twentieth century).

100. Useful summaries of how to approach such literature searches in the social sciences appear in MaryLu C. Rosenthal, *The Fugitive Literature*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS, *supra* note 16, at 85, 86-87, and MaryLu C. Rosenthal, *Bibliographic Retrieval for the Social and Behavioral Scientist*, 22 RES. HIGHER EDUC. 315, 315-17 (1985).

101. A useful beginning point, which reviews some of the primary social science electronic resources, appears in JOHN MONAHAN & LAURENS WALKER, SOCIAL SCIENCE IN LAW app. at 652-54 (6th ed. 2006).

analyst should make the effort to obtain data not formally published as well—for instance, research reported at conference or symposium proceedings, in working papers, or in other works in progress, as well as those reporting null findings.[102] Contacting researchers in the field (e.g., through Listservs, bulletin boards, or other means) can help in such efforts and is a common approach.

Once articles are identified, their bibliographies should be canvassed as well to determine whether the studies cite additional work to be obtained.[103] Finally, a list, table, or appendix should be provided with the review, identifying each of the studies entering into the meta-analysis.

B. Reliability/Quality Judgments

A recurring question in both conducting and evaluating meta-analyses is the quality of the studies being synthesized. As suggested above, some critics object that including studies of "inferior" quality in a meta-analysis in turn reduces the quality of the overall review.[104] Therefore, the argument runs, such studies should be excluded.

At least two related problems exist with this approach, however: the subjectivity of what constitutes "inferior" and the loss of data attendant on excluding studies. There is no question that experiments that lack random assignment, or multiple regression equations that are poorly specified or theorized, represent flawed methodologies. Such flaws, however, do not necessarily justify total exclusion of the study's findings. Rather, careful evaluation and quantification of the study's quality is appropriate.[105] Indeed, exclusion based on perceived flaws is simply an extreme version of such quantification, assigning a weight of 0 to such "inferior" studies and of 1 to those that are included.[106] A more useful approach is to assign a range of weights to the studies, based on criteria determined before conducting the review. This approach is especially useful when such criteria incorporate objective, accepted standards for quality methodology—for instance, random assignment for experimental research. Independent judges can rate the studies according to these criteria and the ratings incorporated into the review.[107] This allows a

102. Legal researchers are becoming accustomed to viewing working papers through various electronic databases such as SSRN, http://www.ssrn.com, and bepress, http://www.bepress.com. Such databases, however, are less common in many domains of empirical social sciences—psychology, sociology, etc.

103. Most social science journals use parenthetical references, with bibliographies, rather than footnote references.

104. See *supra* Part II.F.1 for a discussion of this criticism.

105. *See, e.g.*, Paul M. Wortman, *Judging Research Quality*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS, *supra* note 16, at 97, 98-99, 101-06 (discussing methods to assess quality of research).

106. *See* Rosenthal, *supra* note 31, at 184 (warning that 1.0 weighting system where good studies are included and bad studies are excluded "is often suspect on grounds of weightier bias").

107. This approach has itself been criticized for subjectivity. *See, e.g.*, Sander Greenland & Keith O'Rourke, *On the Bias Produced by Quality Scores in Meta-Analysis, and a Hierarchical View of Proposed Solutions*, 2 BIOSTATISTICS 463, 464, 466-67 (2001) (describing problems associated with quality scores as bias predictors). Specification of the evaluative criteria, however, and ratings by third parties whose reliability can then be statistically analyzed and reported, should eliminate such concern.

reviewer to make use of all available data, giving the best overall sense of the state of a literature. It also permits statistical analysis of the extent to which study quality may have influenced studies' results by analyzing the relationship between these quality ratings and the observed effect sizes for the various studies.

I do not want to imply that all studies in a discipline *must* be included in a meta-analysis; for any number of reasons, studies will likely be found sufficiently flawed or otherwise excludable.[108] What is important, however, is to develop and report an objective means for determining whether to do so and to investigate a possible relationship between those criteria and the studies' observed outcomes.

C. *Coding*

Evaluation of the studies' quality, however, is not the only coding that might be done. The meta-analyst—or judge or practitioner or policy maker—will also be interested in a variety of other variables that might relate to the studies' outcomes and should code for these as well. Numerous descriptive aspects of the studies should be recorded—e.g., publication status, date, sample used, methodology, sample size, equation structure or type, jurisdiction—to develop a list of factors that might be associated with outcome. As noted above, for instance, a meta-analysis of sex differences in perceptions of sexual harassment found that the date of publication was associated with effect size, such that more recent studies found larger sex differences in perceptions.[109] A meta-analysis of studies examining the effect of race on mock juror decisions found an opposite trend, such that studies conducted in the 1970s reported larger racial bias effects than more recent studies.[110] The type of analysis can itself affect outcome: Daniel Pinello, for instance, used statistical technique as a moderator variable in his review of studies examining judicial ideology.[111] He found that multiple regression analyses gave consistently higher effect size estimates than analyses using zero-order correlations.[112]

A meta-analyst will likely also develop theoretically based moderator variables. For instance, in examining whether mock juries might be influenced by pretrial publicity, the type of case might be thought to affect the outcome.[113] Or a meta-analyst reviewing the effects of television violence on antisocial behavior

---

E.g., ROSENTHAL, *supra* note 12, 51-58 (discussing such evaluation and formal methods for assessing reliability).

108. Wortman, *supra* note 105, at 98 (recognizing that issue is not whether to eliminate studies, which is inevitable, but to choose which studies to include or exclude based on quality measurement).

109. *See supra* note 52 and accompanying text.

110. Mitchell et al., *supra* note 9, at 627.

111. Daniel R. Pinello, *Linking Party to Judicial Ideology in American Courts: A Meta-Analysis*, 20 JUST. SYS. J. 219, 237 (1999).

112. *Id.*

113. Nancy Mehrkens Steblay et al., *The Effects of Pretrial Publicity on Juror Verdicts: A Meta-Analytic Review*, 23 LAW & HUM. BEHAV. 219, 224 tbl.2, 224-25, 227 (1999). It seemed to; crimes of murder, sexual abuse, or drugs yielded larger effects of pretrial publicity than did other crimes. *Id.* at 227.

might hypothesize that effects might vary according to the sort of television program (cartoon, news, movie, sports, etc.) someone is exposed to.[114] Based on such theoretical approaches, and on an initial pass through the studies, a reviewer can develop and quantify a list of potential moderator variables to examine.

D. Data Analysis

Once a body of empirical studies has been identified, obtained, rated, and coded into meta-analytic data, those data are ready to be analyzed. The most straightforward synthesis of data can occur with experimental studies—that is, studies that manipulate a specific treatment, randomly assigning the study participants into one or another treatment condition. Quasi-experimental studies, which also look to identify causal connection between a treatment and an outcome measure but typically do not have random assignment to treatment conditions,[115] may also be easily synthesized. The next Parts sketch procedures for combining and comparing experimental studies using two or more conditions, as well as for synthesizing quasi-experimental studies or the common multiple regression studies used in econometric and some empirical legal literature.

1. Two Experimental Conditions

In the simplest such studies, researchers have compared two experimental conditions and have identified whether a statistically significant difference is present between the conditions. Ideally, they have also calculated the strength of that difference—i.e., the effect size. The results of these significance tests and, more important, of the effect size calculation, can be compared and combined with similar data extracted from other studies' results. I provide below general equations for each such procedure.

a. Significance Tests

Again, conducting a significance test yields a "*p*-value," a value between zero and one that describes the likelihood the observed results would have occurred by chance, if there were no true difference between the experimental conditions. And again, significance testing is of less import than effect sizes, both because the *p*-value changes easily depending on the sample size and because of devotion by social scientists and the legal system to the "magical" .05 level of significance.[116] Nevertheless, where focus on significance tests is desired or appropriate, the following approach may be used.

First, the *p*-values must be "one-tailed." In comparing two experimental

114. Paik & Comstock, *supra* note 92, at 528-29, 530 tbl.4. They seemed to; cartoons, for instance, had a higher effect than other sorts of shows. *Id.* at 529.

115. *See* THOMAS D. COOK & DONALD T. CAMPBELL, QUASI-EXPERIMENTATION: DESIGN & ANALYSIS ISSUES FOR FIELD SETTINGS 6 (1979) (explaining purpose and procedures involved in quasi-experiments).

116. See *supra* notes 31-44 and accompanying text for an explanation of the significance of and flaws of *p*-values.

conditions, the researchers might have been interested in whether one group gave results higher or lower than the other—that is, researchers may have expected the conditions to be different, but it may not have been clear a priori in which direction they would differ. If that were the case, researchers would be using a *two*-tailed test, simply looking for significant differences in either direction. Alternatively, researchers may have predicted one condition to yield results significantly higher *or* lower than the other; taking that approach, researchers would use a *one*-tailed test, looking for differences in a particular direction and ignoring differences that were statistically significant, but in the other direction. Using a two-tailed test allows differences in either direction to be identified, but decreases the range in which a difference may be considered statistically significant.[117] That is, if the researcher decides ahead of time that only *p*-values at or less than .05 (i.e., $p <= .05$) will be considered statistically significant, then only differences at the extreme 2.5% values in either direction will qualify. In contrast, with a one-tailed test, any difference in the expected direction up to and including a value of .05 may be considered statistically significant. In both instances the statistical cutoff for significance is the same, but using a two-tailed test requires a smaller range of acceptability, albeit in either direction. For purposes of comparing and combining significance values in meta-analysis, however, all significance levels must be one-tailed.[118]

Second, each *p*-value is converted into its corresponding standard normal deviate, a *Z*-value. This converted *Z*-value represents the point on the standard normal curve that is smaller than *p* percent of the area under the distribution curve. Most introductory statistics textbooks contain tables of the standard normal distribution that permit conversion from *Z*-values to *p*-values and vice versa; Appendix Z provides one here as well.[119]

Again, in converting a *p*-value to its corresponding *Z*, the meta-analyst must keep in mind that a one-tailed *p*-value must be used; and, thus, that *the original researcher's* p*-value must be halved if a two-tailed test were used.* In Appendix Z, the observed *p*-value is located in the left column, and the corresponding *Z*-value is read from the right column. For instance, the observed (or halved) *p*-value might be .05, one-tailed. That value falls halfway between the entries of .0505 and .0495, corresponding to *Z*-values of 1.64 and 1.65, respectively; accordingly, the converted *Z*-value would be 1.645. Of course, the sign of the *Z*-value—positive or negative—depends on the direction of the results and might be different for different studies. Once *Z*-values are obtained for all studies, they can be compared or combined through the equations below.

i. Combining Results of Significance Tests

Again, significance tests from the various studies entering into a meta-

117. For a useful discussion of one- versus two-tailed significance tests, see HOWELL, *supra* note 34, at 92-93, and GEORGE W. SNEDECOR & WILLIAM G. COCHRAN, STATISTICAL METHODS 67-68 (8th ed. 1989).

118. ROSENTHAL, *supra* note 12, at 61.

119. See *infra* Appendix Z for a table converting *p*-values to their corresponding *Z*-values.

analysis examine whether statistically significant differences exist between two experimental conditions. By combining the significance tests from a set of studies, the meta-analyst investigates the *overall* likelihood that that set of *p*-values would have been observed if there were in fact *no* differences between members of those experimental conditions.[120] The meta-analyst calculates the value

$$\frac{\sum Z_i}{\sqrt{K}},$$

where K is the number of studies and Z is the converted Z-value for each observed p-value from Study 1 to Study K. This value is distributed as the standard normal deviate—that is, it is another Z-value. Again, the normal distribution table (Appendix Z) is consulted to find the probability associated with that Z-value—the probability that the observed set of p-values might have been observed by chance, if there were in fact *no* difference between members of those experimental conditions. This time, the Z-value is found in the right column and the corresponding p-value—the consequent overall significance level—is located in the left.

Note that the meta-analyst may have reason to consider some studies more "valid," interesting, or otherwise relevant than others.[121] The studies may better (or worse) approximate the facts of a particular case, be more (or less) rigorous methodologically, or may possess (or lack) any quality rendering them worthy of more (or less) attention. If so, the meta-analyst may want to weight the studies accordingly, by simply assigning numerical weights to the Z-value for each study as follows:

$$Z_{weighted} = \frac{\sum w_i Z_i}{\sqrt{\sum w_i^2}}.$$

Z_i is once more the converted Z-value for each observed p-value, and w_i is the weight assigned by the meta-analyst. Again, that resulting Z-value is tested on the normal distribution table to determine the probability that the observed set of p-values might have been observed by chance, if there were in fact *no* difference between members of those experimental conditions.

ii. Comparing Results of Significance Tests

Given a number of studies comparing two experimental conditions, a meta-analyst might also examine whether those studies' results are significantly different from each other.[122] That is, the heterogeneity of the significance tests might be of interest, as one way of determining whether further exploration of

120. *E.g.*, ROSENTHAL, *supra* note 12, at 85 (going through analysis for combining *p*-values for two example sets of data).

121. See *supra* Part III.C for issues regarding coding based on quality or various other factors.

122. Robert Rosenthal & Donald B. Rubin, *Comparing Significance Levels of Independent Studies*, 86 PSYCHOL. BULL. 1165, 1166-67 (1979).

moderator variables is appropriate.

Again, the meta-analyst ensures that all *p*-values are one-tailed and that the valence (positive or negative) accurately reflects the direction of the findings. For *K* studies, each *p*-value is converted to a *Z*-value, as described above,[123] and the mean of those *Z*-values is calculated. The mean is then subtracted from each value and squared; those values are then summed:

$$\sum\left(Z_i - \overline{Z}\right)^2 .$$

The resulting value is distributed as chi-square (χ^2), with *K*-1 degrees of freedom, i.e., the number of studies (*K*) minus 1.[124] That is, a table of the "critical values" for χ^2 is consulted (e.g., Appendix X, *infra*). The appropriate degrees of freedom (*df*) is calculated. The observed *Z*-value is then compared to the values in the appropriate row of the table; the value is statistically significant at a level determined by the column whose critical value is *less than* the observed *Z*-value. For instance, at 20 degrees of freedom, an observed *Z*-value of 32.00 is statistically significant at the .05 level because the critical value in that column, 31.41, is lower. It is not statistically significant at the .025 level, however, because the critical value in that column, 34.17, is higher. A statistically significant result indicates that there is heterogeneity among the significance tests in the studies and that further exploration is likely warranted.

b. Effect Sizes

More important than comparing and combining significance tests is examining the effect sizes yielded by the individual studies entering into the meta-analysis. An "effect size" reflects the strength or magnitude of the relationship between two variables,[125] though there are a number of different statistics that can be used to describe that magnitude. The two most common such statistics are *r* and *Cohen's d*; for several reasons, however—ease of use, increased familiarity to nonstatisticians, and ease of conversion into measures of practical effect[126]—the former, *r*, is emphasized here.[127]

The effect size *r* is often simply the correlation between the two variables of

123. *See supra* note 119 and accompanying text.

124. A table giving critical values for testing chi-squares is provided *infra* in Appendix X.

125. ROSENTHAL, *supra* note 12, at 14.

126. Blumenthal, *supra* note 51, at 39; Orr & Guthrie, *supra* note 4, at 613.

127. *Cohen's d* is commonly used in experimental research, however, and a meta-analyst might come across studies using either *r* or *d*. If so, one can be calculated from the other according to the following equations. To go from *r* to *d*:

$$d = \frac{2r}{\sqrt{1-r^2}} .$$

To go from *Cohen's d* to *r*,

$$r = \frac{d}{\sqrt{d^2+4}} ,$$

so long as the sample sizes of the two groups being compared to calculate *d* are equal or about equal. If not, a more generalized form of the equation is available. ROSENTHAL, *supra* note 12, at 19.

interest. However, when the original researcher reports instead the significance test addressing the differences between the two groups—for instance, a *t*-test or an *F*-test for continuous data—then *r* can be calculated from that statistic. If the researcher reports the *t*-value describing the difference between the two groups, *r* can be calculated as follows:

$$r = \sqrt{\frac{t^2}{t^2 + df}} \text{ [Equation R}^{128}\text{]},$$

where *df* is the degrees of freedom for the significance test (n_1+n_2-2, with n_i being the sample size for each experimental group or condition). Because the statistic *F* is simply the square of a *t*-value, a similar equation may be used when the original researcher reports an *F*-test from an analysis of variance or ANOVA:

$$r = \sqrt{\frac{F}{F + df}},$$

where *df* is the degrees of freedom associated with the error term from the analysis of variance (for instance, a researcher will report the *F*-value as F(1,100), indicating that the relevant *df* to use for the effect size calculation is 100).

Importantly, this effect size calculation for *F* is far less useful when the numerator of the *F*-test (i.e., the first term in the parentheses) is greater than one—that is, when the original researcher is reporting an *F*-test on more than two groups (for instance, a researcher may report the *F*-value as F(2,100), indicating that she was comparing differences across *three* groups[129] and approximately 100 subjects). This is because such a comparison does not specifically test what difference may exist between any two of the groups, but rather whether some difference at all may exist somewhere among the groups.[130] Knowing the latter is less helpful practically than determining the former.

In some instances, the original data may not have been "continuous"—e.g., answers on an opinion poll describing agreement with a given statement on a one to five scale—but rather "categorical"—e.g., the proportion of cases won or the number of people responding to a dichotomous question. In the latter instance, a primary researcher will report the results of a χ^2 (chi-square) test, testing the relationship between two variables of interest. For instance, a study might compare whether men or women view the same workplace behavior as sexual harassment. The researcher would tally the number of men and women reporting yes and no and test whether there is a relationship between the two variables *sex* and *perception of sexual harassment.* A statistically significant result on the original researcher's χ^2 test would suggest a relationship between the two

128. I refer to this equation *infra* in the text following note 143 and the text accompanying note 145.

129. As a methodological matter, the degrees of freedom in the numerator is always one fewer than the number of groups being compared.

130. *E.g.*, ROSENTHAL, *supra* note 12, at 13 (commenting that *F*-test with *df* greater than one in numerator results in quantitative answers that are "hopelessly imprecise"). This broader test is called a "diffuse" test; the test with one *df* in the numerator is a "focused" test.

variables, i.e., that whether a person views such behavior as sexual harassment depends on that person's sex. For a meta-analyst to then calculate an effect size for such categorical data that can be combined with or compared to effect size *r*'s, the following equation is used:

$$\phi = \sqrt{\frac{\chi^2(1)}{N}},$$

where $\chi^2(1)$ is the result of the χ^2 test at one degree of freedom and N is the total sample size. The statistic Φ (phi) can then be interpreted as an *r*.

As in the discussion above, however, note that diffuse χ^2 tests, comparing more than two categories (e.g., answers from Republicans, Democrats, and Independents), with no further information, are less than helpful—finding statistically significant results may demonstrate that *some* difference appeared among those groups but would not identify which group was different from the others.[131] Thus, a diffuse χ^2 test is not amenable to the equation above and is why the equation emphasizes that only one degree of freedom is present.

i. Combining Effect Sizes

Perhaps the most important calculations in a meta-analysis—and among the most straightforward—involve combining the effect size *r*'s that reflect the findings from each study. That is, because of the importance of effect sizes relative to significance tests as reflecting the outcome of empirical studies, synthesis and subsequent analysis of the *r*'s gives the best summary of a body of empirical work.

Once the effect size *r*'s are calculated as outlined above, they can be combined and compared. Correlational analysis can also be conducted to determine whether relationships with various moderator variables exist. Before undertaking such analyses, however, the *r*'s must be transformed for statistical purposes—when untransformed, a skew can emerge when *r*'s become large (i.e., as they get further from zero). Accordingly, statisticians have developed a transformation that is essentially unbiased and should be used for calculations. That is, for each effect size *r* that is obtained, a transformed effect size estimate—known as "Fisher's z_r"—should be calculated and any analyses should be conducted on those values.[132] To transform an effect size *r* to a Fisher's z_r, the following equation is used:

$$z_r = \frac{\log_e\left(\frac{1+r}{1-r}\right)}{2}.$$[133]

A useful shorthand table appears in Appendix RZR, where values for Fisher's z_r

131. ROSENTHAL, *supra* note 12, at 15.

132. *E.g.*, MARK W. LIPSEY & DAVID B. WILSON, PRACTICAL META-ANALYSIS 63 (2001) (discussing need for this statistical transformation).

133. Note that this is the natural logarithm (i.e., with *e* as the base, rather than with 10 as the base).

can be found for values of r to two decimal points.

For combining effect size r's, the following equation is used to determine the average:

$$\overline{z}_r = \frac{\sum z_r}{K},$$

where K is the total number of entries.[134] Again, a meta-analyst might deem it appropriate to weight the effect sizes based on some criteria, whether sample size, methodological rigor, similarity to a case at hand, or other factor. If so, the weighted z_r may be calculated as follows:

$$\overline{z}_r = \sum \frac{w_i z_{r_i}}{w_i},$$

where w is the weight assigned a priori to each study i.

Calculation of, for instance, the mean, median, and other measures of central tendency should be conducted on the converted Fisher's z_r's, with the resulting value then being converted back to the more understandable effect size r.[135] The conversion may be done either by consulting the shorthand table in Appendix ZRR or by applying the equation

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1};$$

e can be found on virtually any statistical package or scientific calculator and equals approximately 2.718.[136]

ii. Comparing Effect Sizes

The process for comparing the heterogeneity of effect sizes—for

134. Note that the summation is of the converted Fisher z_r's, rather than calculating the mean r and then converting that r into a Fisher z_r.

135. *E.g.*, LIPSEY & WILSON, *supra* note 132, at 64 (noting that retransformation); ROSENTHAL, *supra* note 12, at 21 (suggesting that although z_r makes a serviceable effect size estimate, r is more easily interpreted).

136. Some meta-analysts recommend further adjustments, in order to "correct" for a variety of potential sorts of unreliability in the data, such as variation in range in the dependent and independent variables, unreliability in those variables, differences in their sample size, etc. *E.g.*, HUNTER ET AL., *supra* note 21, at 35-92 (describing several ways of correcting statistical bias). This approach certainly helps focus the data and demonstrate "what effect size we might expect to find in the best of all possible worlds." ROSENTHAL, *supra* note 12, at 24. Nevertheless, it is not clear that this is necessarily what a meta-analyst—or a judge or policy maker—in fact wishes to learn. *Id.* at 25. What is at issue in a synthetic review of an empirical literature is what we know about existing research, not what might be the case if all variables were "perfectly measured, perfectly valid, perfectly continuous, and perfectly unrestricted in range." *Id.* More specifically,

> focusing on unadjusted effect size estimates allows evaluation of the existing literature as it actually stands. When 'corrections' are not used, analysis can reflect what researchers actually found, rather than what they might have found. Moreover, it can help identify potential moderators of the effects at hand, examining what factors may have led to any biases observed.

Blumenthal, *supra* note 51, at 39 (citations omitted).

determining whether observed effect sizes are statistically significantly different from each other—mirrors that for comparing differences among significance levels described in Part III.D.1.a.ii. Again, each effect size r is obtained and converted into a Fisher's z_r. The degrees of freedom associated with each r is also calculated as $(N-3)$, where N is the number of sampling units on which each r is based.[137]

The following value is then tested on a chi-square distribution table as explained above; that is, the term

$$\sum (N_i - 3)\left(z_{r_i} - \overline{z_r}\right)^2$$

is distributed as chi-square.[138] Note, however, that the mean z_r value, $\overline{z_r}$, is not simply the unadjusted mean, but rather a mean weighted by the sample size or degrees of freedom associated with each effect size r. More specifically, it is

$$\overline{z_r} = \frac{\sum (N_i - 3) z_{r_i}}{\sum (N_i - 3)}.$$

A statistically significant result indicates that there is heterogeneity among the effect sizes in the studies, warranting further exploration to evaluate the effect of moderator variables.

2. More Than Two Experimental Conditions

As suggested above, there is a tendency to apply diffuse tests (comparing more than two conditions) in an effort to make specific inferences about differences among multiple conditions, when such tests in fact tell us little other than that some difference might exist.[139] A better way of investigating multiple experimental conditions involves the use of *contrast analysis*.[140] Using contrasts, an experimenter uses existing theory to predict a pattern of findings (e.g., Republicans voting for X more often than Democrats, with Independents in between; or twelve-person juries deliberating for longer than eight-person juries, who in turn deliberate longer than six-person juries). Based on such predictions, the experimenter assigns weights to each condition then statistically analyzes how closely the weighted predictions match the observed data.

Although such contrast analysis is a more powerful and precise statistical tool for detecting patterns and differences among multiple experimental conditions, it is rarely used in the context of empirical legal research. Nevertheless, it should be. Because of its current rarity, however, I will not elaborate on tools for meta-analyzing studies using contrast analysis, but simply

137. ROSENTHAL, *supra* note 12, at 73.

138. See *infra* Appendix X for a list of the critical values of x^2.

139. See *supra* notes 132-38 and accompanying text for how effect sizes affect the comparison of those effect sizes.

140. *See generally* ROBERT ROSENTHAL & RALPH L. ROSNOW, CONTRAST ANALYSIS: FOCUSED COMPARISONS IN THE ANALYSIS OF VARIANCE (1985) (providing in-depth description of value of contrast analysis); ROBERT ROSENTHAL ET AL., CONTRASTS AND EFFECT SIZES IN BEHAVIORAL RESEARCH: A CORRELATIONAL APPROACH (2000) (looking at wider and more useful application of contrast analysis by introducing correlation effect size estimates).

identify other resources for doing so.[141]

3. Regressions

Much of the empirical legal literature involves econometric and other regression analysis (arguably, there is a tendency to equate "empirical legal studies" with econometric legal scholarship, with a substantially less visible role for experimental research). Despite some objections, and although it is not as straightforward as meta-analyzing experimental studies, there are a number of ways to synthesize this sort of empirical literature as well.

There are multiple ways to derive effect size estimates from such multiple regression studies. In the first approach, possessing the virtue of simplicity, the meta-analyst might simply treat the *standardized* regression coefficients (betas) as effect sizes and combine them in the ordinary way described above.[142] To repeat, this combination must address the *standardized* coefficients, as opposed to the simple slopes, whose units are explicitly dependent on the context of the specific study.[143] When enough information is provided in the initial study, the meta-analyst might use the second approach, converting regression coefficients to effect sizes through Equation R, above. More specifically, some equations provide *t*-values for the relevant coefficient. That *t*-value can be converted to an effect size *r* (the partial correlation), using as the degrees of freedom the value *N-p*-1, where *N* is the total sample size associated with the particular predictor and *p* is the number of predictors in the equation (other than the intercept).[144] When a *t*-value is not explicitly provided, however, one may be calculated if the standard error of the coefficient is provided:

$$t = \frac{b}{se_b},$$

where *b* is the regression coefficient and se_b is its standard error. The difficulty of this approach is the relative infrequency with which the total sample size (*N*) is provided. Of course, if the *N* is not given, then by examining the original study the meta-analyst might be able to determine and obtain the data set used, thus

141. *E.g.*, ROSENTHAL, *supra* note 12, at 79-81 (reviewing meta-analysis of contrast studies).

142. As Rosenthal and DiMatteo point out:

> The standardized *beta* from a multiple regression, as well as a partial correlation, can be used as effect size estimates, but it must be remembered that these represent the relationship between the independent and the dependent variable controlling for other factors (and the meta-analyst might want separately to combine *r*'s and partial *r*'s/standardized *betas*).

Rosenthal & DiMatteo, *supra* note 31, at 72.

143. JOHN E. HUNTER & FRANK L. SCHMIDT, METHODS OF META-ANALYSIS: CORRECTING ERROR AND BIAS IN RESEARCH FINDINGS 192-95 (2d ed. 2004) [hereinafter HUNTER & SCHMIDT (2d ed.)].

144. The partial correlation derived from the standardized regression coefficient is more appropriate to use than the simple correlation, which does not take into account the other variables in the multiple regression equation. Chris Doucouliagos & Patrice Laroche, *Unions and Productivity Growth: A Meta-Analytic Review*, *in* THE DETERMINANTS OF THE INCIDENCE AND THE EFFECTS OF PARTICIPATORY ORGANIZATIONS 57, 77 n.5 (Takao Kato & Jeffrey Pliskin eds., 2003).

deriving the sample size investigated and thus the N to use in Equation R.[145]

One important caveat must be placed on the synthesis of multiple regression studies. Regression estimates vary, sometimes substantially, depending on what other variables are included in an equation.[146] Such variability will affect the magnitude of the coefficient and thus, of course, its statistical significance. Some scholars suggest that this variability casts doubt on the usefulness of testing regression coefficients,[147] and on the usefulness of including such estimates as data for meta-analyses.[148]

Such criticism, however, is ultimately unpersuasive.[149] First, the variability in predictive value in a regression equation as a function of what predictors are included is little different conceptually from the conventional issues in experimental research of identifying and controlling for hidden or third-variable influences. That is, experimental research examining the correlation between variables X and Y may or may not control for variable Z that in fact influences that relationship. Identifying and controlling for Z will affect the strength of the researchers' findings, but whether a study does so need not serve as the basis for including or excluding it in a qualitative or quantitative review.[150] Each of the studies is returning an estimate—to a more or less sophisticated degree—of the relationship between variables X and Y, and may therefore warrant being included in a synthesis.

Second, it is plausible that any set of predictor variables used in different studies will have some theoretical justification, though the individual researchers might disagree about what modeling strategy and what variables are appropriate.

---

145. As a third method, one might combine the covariance matrices of the different predictor variables. Francesca Dominici et al., *Combining Information from Related Regressions*, 2 J. AGRIC. BIOLOGICAL & ENVTL. STAT. 313, 316-19 (1997). That information is even more rarely provided in the empirical legal literature.

146. JOHN E. HUNTER & FRANK L. SCHMIDT, METHODS OF META-ANALYSIS: CORRECTING ERROR AND BIAS IN RESEARCH FINDINGS 502 (1990) ("[B]eta weights are relative to the set of predictors considered and will only replicate across studies if the exact set of predictors is considered in each. If any predictor is added or subtracted from one study to the next, then the beta weights for all variables may change."). In the second edition of their text, however, Professors Hunter and Schmidt imply that using standardized coefficients (i.e., beta weights) may ameliorate this concern, as suggested in the text above. HUNTER & SCHMIDT (2d ed.), *supra* note 143, at 194. Their focus, however, is only on simple (i.e., bivariate) regression, not multiple regression. *Id.*

147. *Cf., e.g.*, HOWELL, *supra* note 34, at 494 (noting that "a test on a variable is done in the context of all other variables in the equation," and thus, its contribution to predicting outcome may change depending on what other variables are included).

148. *E.g.*, HUNTER & SCHMIDT (2d ed.), *supra* note 143, at 475 (suggesting that "regression weights are typically not suitable for cumulation"). *But cf.* Christos Doucouliagos & Patrice Laroche, *What Do Unions Do to Productivity? A Meta-Analysis*, 42 INDUS. REL. 650, 658 (2003) (explaining usefulness of meta-regression analysis and using it despite challenges).

149. Note too that even traditional narrative reviews will be subject to such criticism. *See* Doucouliagos & Laroche, *supra* note 148, at 654-55 (criticizing traditional qualitative reviews for being overly subjective and speculative and unable to scientifically assess specification differences).

150. Analogously, multiple regression studies may be better or worse at specifying the "correct" or consistent model or at including the "correct" or relevant or consistent predictors. Each, however, is testing the relationship between predictors and an outcome, and the studies are not per se incomparable.

Accordingly, there is not necessarily a clear a priori basis for saying that one set of variables is "correct," and therefore for dismissing certain studies and retaining others.[151] The studies may be rated by separate judges, and the studies' contributions weighted by those ratings[152]—with the presence or absence of theoretically important variables affecting those ratings—but inconsistency across studies in which predictors are used does not necessarily mean synthesis of those studies is inappropriate.

Third, meta-analysis is in fact the more useful approach to investigating such nonidentical studies, through moderator analysis.[153] Suppose a meta-analyst identifies five multiple regression studies predicting outcome variable *Y*. Studies 1, 2, and 3 use the set of predictor variables *A*, *B*, and *C*. Study 4 uses predictor variables *B*, *C*, and *D*, while Study 5 expands on Study 4 and uses variables *B*, *C*, *D*, and *E*. Critics applying the stricter approach sketched above would suggest that only Studies 1, 2, and 3 may be synthesized, as those are the only directly comparable studies.[154] Clearly, though, as in any synthesis, the presence or absence of a predictor variable, or set of predictors, can become a variable itself in the meta-analysis. The meta-analyst might dummy code for the presence of a variable (or set) and test the influence and statistical significance of that dummy variable. Thus, Studies 4 and 5, which include variable *D*, might be compared to Studies 1, 2, and 3, which do not. A significant difference between those sets of studies would tell something about the influence of variable *D* on the research findings (and on its impact on other predictor variables). Of course, depending on the number of studies, smaller meta-analyses might be done of studies that *do* include exactly the same variables (such as Studies 1, 2, and 3 here).

Accordingly, with appropriate consideration of model specification and quality, variation in the (sets of) predictors used, type and amount of information provided, and other factors, the procedures identified above can be used to summarize and synthesize effect sizes from multiple regression studies. A meta-analyst can also examine all of the types of moderators sketched earlier, in addition to dummy coding for different types and sets of predictor variables. The estimates derived from these studies could also be combined with other studies in the meta-analysis that obtained effect sizes in other ways (e.g., experimental studies).[155]

---

151. Compare *supra* notes 83-84.

152. See *supra* Part III.B for an explanation of this and other procedures that may be used to address the reliability or quality of the synthesized studies.

153. See *supra* Part III.C for an explanation of the possibilities for coding for various moderator variables.

154. *E.g.*, HUNTER & SCHMIDT (2d ed.), *supra* note 143, at 475 (providing example of how to choose studies with common variables to compare). Professors Hunter and Schmidt would also consider synthesizing studies that might include the same predictor variables (alone or as a subset), so long as the study published the full set of intercorrelations among the predictors. *Id.* This is rare, especially in legal literature.

155. There are at least three other alternatives for a researcher seeking to synthesize a multiple regression literature. First, one might use "meta-regression analysis," an approach expressly designed to summarize multiple regression studies in empirical economics. T.D. Stanley, *Wheat from Chaff: Meta-Analysis As Quantitative Literature Review*, 15 J. ECON. PERSP. 131, 131 (2001); T.D. Stanley &

E. *Data Reporting*

1. Summary Statistics and the Binomial Effect Size Display

The final step in a meta-analysis is to report the results of the various analyses. First, as a descriptive matter, once all the effect sizes are calculated, a stem-and-leaf plot can be presented in order to give a pictorial sense of the range of observations.[156] A table giving summary information about those effect sizes should be presented as well, including the mean, the weighted mean (if weights were used), standard deviation, range, and number of studies entering into those means.

Once an overall summary effect size is calculated, however, how does a meta-analyst (or attorney or court or policy maker) know—and convey—how "important" or "significant" it is in practical terms? This is especially an issue when the mean effect size might be relatively "low," explaining little of the overall variance in observations. Consumers of such a review might be tempted to dismiss small effect sizes as unimportant, or, again, to dismiss as illusory even large effects that do not reach conventional significance levels.

Meta-analysts have developed a simple, intuitive method for presenting

---

Stephen B. Jarrell, *Meta-Regression Analysis: A Quantitative Method of Literature Surveys*, 3 J. ECON. SURVEYS 54 (1989), *reprinted in* 19 J. ECON. SURVEYS 299, 301 (2005) ("Simply stated, to review empirical economic literature, one must summarize regression results."). Second, one might apply a variation of hierarchical modeling and data augmentation techniques. *E.g.*, Dominici et al., *supra* note 145, at 314 (proposing "combination of hierarchical modeling and data augmentation" to deal with combining multiple regressions). This approach is useful in addressing issues that arise when studies use different sets of predictors or have missing data. Nonetheless, substantially more information from the primary studies is necessary, often prohibitively so. *Id.* at 331 ("[W]e assumed throughout that the study's means and covariance matrices are available for analysis. Meta-analysis of regression studies requires different approaches when more limited information, such as significance test results, is reported."). Finally, an alternative to explicit synthesis that nevertheless is useful in drawing causal inferences from the sorts of large databases that econometric analysis often utilizes is "propensity score analysis." This approach is a distant cousin—though arguably a more elegant and precise relative—of comparing studies with different sets of predictors. It is also quite similar to the "nonparametric matching" approach recently used by Epstein and colleagues. *See* Lee Epstein et al., *The Supreme Court During Crisis: How War Affects Only Non-War Cases*, 80 N.Y.U. L. REV. 1, 65-69 (2005) (describing their procedure, "which uses the insights of random assignment to draw causal inferences in observational studies, while decreasing the role of onerous assumptions of conventional parametric estimates"). More specifically, propensity score analysis controls for different sets of naturally occurring background characteristics that are likely not controlled for in different studies, "reducing the entire collection of background characteristics to a single composite characteristic that appropriately summarizes the collection." Donald B. Rubin, *Estimating Causal Effects from Large Data Sets Using Propensity Scores*, 127 ANNALS INTERNAL MED. 757, 757 (1997); *see also* Ralph B. D'Agostino, Jr. & Donald B. Rubin, *Estimating and Using Propensity Scores with Partially Missing Data*, 95 J. AM. STAT. ASS'N 749, 749 (2000) (explaining that "[p]ropensity scores are a one-dimensional summary of multidimensional covariates, X, such that when the propensity scores are balanced across the treatment and control groups, the distribution of all the covariates, X, are balanced in expectation across the two groups"); Paul R. Rosenbaum & Donald B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, 70 BIOMETRIKA 41, 41 (1983) (explaining that "propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates").

156. *E.g.*, Blumenthal, *supra* note 51, at 40 tbls.2 & 3 (illustrating stem-and-leaf plots of effect sizes).

such effect size findings that can help avoid such potentially unwarranted dismissal: the binomial effect size display ("BESD").[157] The BESD illustrates in tabular form the impact of an effect in real terms. It focuses on the change in success rate—defined in terms of the relationship in question—attributable to the particular effect. To do so, a table is initially constructed showing success and failure rates (in whatever terms might be relevant) as though no difference existed—for instance, a 50/50 split. The observed effect size r is then divided by two, and the resulting sum is added to (or subtracted from, if the observed effect were negative) the default "success rate"; the opposite is done to the default "failure rate." A table with the resulting values is then presented.

As a straightforward example, suppose the relationship in question (the mean effect size across a number of medical studies) was $r = .30$, describing the effectiveness of a particular treatment being challenged in court. The percent of the variance explained by that treatment (i.e., the contribution it makes to an improvement in health) is $r^2 = .09$—less than ten percent, superficially unimpressive. When that effect is described in tabular form through the BESD, however, it appears more substantial. First, the table is constructed as though no difference existed, with a 50/50 split between conditions. Then the observed r, .30, is divided by two (yielding .15), is added to the success rate, and is subtracted from the failure rate. The resulting table is as follows:

Treatment Result[158]

	Improvement	No Improvement	Total
Treatment	65	35	100
Control	35	65	100
Total	100	100	200

Clearly, even this "small" overall effect, explaining "only" nine percent of the variance in effect, has a substantial impact on the improvement rate of those experiencing the treatment. Indeed, far smaller effects have been seen as of substantial practical importance,[159] and the BESD can help demonstrate why. Obviously, the same sort of presentation can help illustrate the importance of large effects that do not reach conventional levels of statistical significance.

2. Moderator Variables

After presenting descriptive summaries of the effect size information, the meta-analyst should present information about the moderator variables and their effects. For instance, a table might be given listing the various moderator variables, the correlation between effect size and each variable and the significance level of the correlation, and a confidence interval around that

157. *See generally* Robert Rosenthal & Donald B. Rubin, *A Simple, General Purpose Display of Magnitude of Experimental Effect*, 74 J. EDUC. PSYCHOL. 166 (1982) (describing BESD, which displays change in success rate attributable to certain treatment procedures).

158. Table patterned after ROSENTHAL, *supra* note 12, at 134 tbl.7.2.

159. *See, e.g.*, ROSENTHAL, *supra* note 12, at 134-36 & tbl.7.5 (describing propranolol study with effect accounting for one-fifth of one percent of variance).

estimate.[160] This encourages discussion, both theoretical and practical, of factors that may or may not have influenced the studies' findings, and the implications of such findings.

Finally, a useful approach to further investigating the moderator variables takes into account that they might not only be related to the observed effect size outcomes, but also be related to each other.[161] Detailed procedures for taking this into account have been developed,[162] but a relatively straightforward approach is meta-regression analysis ("MRA"), mentioned above.[163] Here, the moderator variables are included in a regression analysis as independent variables predicting the effect sizes observed in the set of meta-analysis studies. MRA allows the meta-analyst to parcel out the effects of potentially correlated moderator variables and provides another means of finding which ones might have an identifiable influence on study outcome.

CONCLUSION

Empirical legal scholarship is mushrooming, and courts' and scholars' use of empirical research from other disciplines is increasing as well. As such empirical work develops, it is essential to have periodic syntheses of the various bodies of research, not only for understanding the state of knowledge in a research field, but also to be able to present such findings to a court or to policy makers, and to develop avenues, both substantive and methodological, for further research. Quantitative review—meta-analysis—serves these goals better than traditional reviews by (1) identifying and synthesizing a larger set of the research in question; (2) evaluating each of the studies—that is, each of the "data points"—that enter into the review; (3) facilitating the comparison or juxtaposition of different studies and the identification of "aberrant" or outlying studies; and (4) identifying "moderator" variables that both can illustrate why a certain study or set of studies resulted a certain way and can generate further hypotheses to investigate.[164] Empirical legal scholars should be familiar with, and make use of, the meta-analytic approach. My goal here is to increase that familiarity and to encourage that use.

---

160. *E.g.*, Blumenthal, *supra* note 51, at 43 tbl.6 (summarizing moderator effects in study on reasonable woman standard for sexual harassment).

161. *See* Lipsey, *supra* note 46, at 69-70 (investigating hazards and complexities of interpreting moderator variables especially when they are related to one another).

162. *E.g.*, Stephen W. Raudenbush, *Random Effects Models*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS, *supra* note 16, at 301, 302 (illustrating random effects approach and weighted least squares regression approach to deal with issues surrounding moderators).

Note that Raudenbush's chapter also highlights an important difference in types of meta-analyses, those using *random effects* models and those using *fixed effects* models. The difference is important, reflecting issues of generalizability and other inferences. *See, e.g.*, Blumenthal, *supra* note 51, at 47 (noting some differences). A good discussion of the differences and reasons for choosing one approach over the other appears in Cooper and Hedges, *supra* note 93, at 526-27.

163. *See* Stanley, *supra* note 155, at 131-32 (discussing potential of meta-analysis to summarize, evaluate, and analyze empirical economic research).

164. Blumenthal, *supra* note 1, at 45.

Appendix Z
Converting *p*-value to Corresponding *Z*-value[165]

p-value	*Z*-value	*p*-value	*Z*-value
.5000	.00	.3669	.34
.4960	.01	.3632	.35
.4920	.02	.3594	.36
.4880	.03	.3557	.37
.4840	.04	.3520	.38
.4801	.05	.3483	.39
.4761	.06	.3446	.40
.4721	.07	.3409	.41
.4681	.08	.3372	.42
.4641	.09	.3336	.43
.4602	.10	.3300	.44
.4562	.11	.3264	.45
.4522	.12	.3228	.46
.4483	.13	.3192	.47
.4443	.14	.3156	.48
.4404	.15	.3121	.49
.4364	.16	.3085	.50
.4325	.17	.3050	.51
.4286	.18	.3015	.52
.4247	.19	.2981	.53
.4207	.20	.2946	.54
.4168	.21	.2912	.55
.4129	.22	.2877	.56
.4090	.23	.2843	.57
.4052	.24	.2810	.58
.4013	.25	.2776	.59
.3974	.26	.2743	.60
.3936	.27	.2709	.61
.3897	.28	.2676	.62
.3859	.29	.2643	.63
.3821	.30	.2611	.64
.3783	.31	.2578	.65
.3745	.32	.2546	.66
.3707	.33	.2514	.67

165. Recall that the *p*-value must be one-tailed, either in the original study or as a result of halving the researcher's value. See *supra* notes 117-18 and accompanying text for a discussion of the requirement that the *p*-value must be one-tailed.

p-value	*Z*-value	*p*-value	*Z*-value
.2483	.68	.1539	1.02
.2451	.69	.1515	1.03
.2420	.70	.1492	1.04
.2389	.71	.1469	1.05
.2358	.72	.1446	1.06
.2327	.73	.1423	1.07
.2296	.74	.1401	1.08
.2266	.75	.1379	1.09
.2236	.76	.1357	1.10
.2206	.77	.1335	1.11
.2177	.78	.1314	1.12
.2148	.79	.1292	1.13
.2119	.80	.1271	1.14
.2090	.81	.1251	1.15
.2061	.82	.1230	1.16
.2033	.83	.1210	1.17
.2005	.84	.1190	1.18
.1977	.85	.1170	1.19
.1949	.86	.1151	1.20
.1922	.87	.1131	1.21
.1894	.88	.1112	1.22
.1867	.89	.1093	1.23
.1841	.90	.1075	1.24
.1814	.91	.1056	1.25
.1788	.92	.1038	1.26
.1762	.93	.1020	1.27
.1736	.94	.1003	1.28
.1711	.95	.0985	1.29
.1685	.96	.0968	1.30
.1660	.97	.0951	1.31
.1635	.98	.0934	1.32
.1611	.99	.0918	1.33
.1587	1.00	.0901	1.34
.1562	1.01	.0885	1.35

p-value	*Z*-value	*p*-value	*Z*-value
.0869	1.36	.0446	1.70
.0853	1.37	.0436	1.71
.0838	1.38	.0427	1.72
.0823	1.39	.0418	1.73
.0808	1.40	.0409	1.74
.0793	1.41	.0401	1.75
.0778	1.42	.0392	1.76
.0764	1.43	.0384	1.77
.0749	1.44	.0375	1.78
.0735	1.45	.0367	1.79
.0721	1.46	.0359	1.80
.0708	1.47	.0351	1.81
.0694	1.48	.0344	1.82
.0681	1.49	.0336	1.83
.0668	1.50	.0329	1.84
.0655	1.51	.0322	1.85
.0643	1.52	.0314	1.86
.0630	1.53	.0307	1.87
.0618	1.54	.0301	1.88
.0606	1.55	.0294	1.89
.0594	1.56	.0287	1.90
.0582	1.57	.0281	1.91
.0571	1.58	.0274	1.92
.0559	1.59	.0268	1.93
.0548	1.60	.0262	1.94
.0537	1.61	.0256	1.95
.0526	1.62	.0250	1.96
.0516	1.63	.0244	1.97
.0505	1.64	.0239	1.98
.0495	1.65	.0233	1.99
.0485	1.66	.0228	2.00
.0475	1.67	.0222	2.01
.0465	1.68	.0217	2.02
.0455	1.69	.0212	2.03

p-value	*Z*-value	*p*-value	*Z*-value
.0207	2.04	.0087	2.38
.0202	2.05	.0084	2.39
.0197	2.06	.0082	2.40
.0192	2.07	.0080	2.41
.0188	2.08	.0078	2.42
.0183	2.09	.0075	2.43
.0179	2.10	.0073	2.44
.0174	2.11	.0071	2.45
.0170	2.12	.0069	2.46
.0166	2.13	.0068	2.47
.0162	2.14	.0066	2.48
.0158	2.15	.0064	2.49
.0154	2.16	.0062	2.50
.0150	2.17	.0060	2.51
.0146	2.18	.0059	2.52
.0143	2.19	.0057	2.53
.0139	2.20	.0055	2.54
.0136	2.21	.0054	2.55
.0132	2.22	.0052	2.56
.0129	2.23	.0051	2.57
.0125	2.24	.0049	2.58
.0122	2.25	.0048	2.59
.0119	2.26	.0047	2.60
.0116	2.27	.0045	2.61
.0113	2.28	.0044	2.62
.0110	2.29	.0043	2.63
.0107	2.30	.0041	2.64
.0104	2.31	.0040	2.65
.0102	2.32	.0039	2.66
.0099	2.33	.0038	2.67
.0096	2.34	.0037	2.68
.0094	2.35	.0036	2.69
.0091	2.36	.0035	2.70
.0089	2.37	.0034	2.71

p-value	*Z*-value
.0033	2.72
.0032	2.73
.0031	2.74
.0030	2.75
.0029	2.76
.0028	2.77
.0027	2.78
.0026	2.79
.0026	2.80
.0025	2.81
.0024	2.82
.0023	2.83
.0023	2.84
.0022	2.85
.0021	2.86
.0021	2.87
.0020	2.88
.0019	2.89
.0019	2.90
.0018	2.91
.0018	2.92
.0017	2.93
.0016	2.94
.0016	2.95
.0015	2.96
.0015	2.97
.0014	2.98
.0014	2.99
.0013	3.00
.0006	3.25
.0002	3.50
.0001	3.75
.0000	4.00

Appendix X: Critical Values of χ^2

df	.90	.75	.50	.250	.10	.05	.025	.01	.005
1	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
10	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.75
12	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.21	28.30
13	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	7.79	10.17	13.34	17.12	21.06	23.69	26.12	29.14	31.31
15	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.15
19	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.56	40.00
21	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	14.04	17.24	21.34	26.04	30.81	33.93	36.78	40.29	42.80
23	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	15.66	19.04	23.34	28.24	33.20	36.42	39.37	42.98	45.56
25	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.32	46.93
26	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	18.11	21.75	26.34	31.53	36.74	40.11	43.20	46.96	49.64
28	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	29.06	33.67	39.34	45.61	51.80	55.75	59.34	63.71	66.80
50	37.69	42.95	49.34	56.33	63.16	67.50	71.42	76.17	79.52
60	46.46	52.30	59.34	66.98	74.39	79.08	83.30	88.40	91.98
70	55.33	61.70	69.34	77.57	85.52	90.53	95.03	100.44	104.24
80	64.28	71.15	79.34	88.13	96.57	101.88	106.63	112.34	116.35
90	73.29	80.63	89.33	98.65	107.56	113.14	118.14	124.13	128.32
100	82.36	90.14	99.33	109.14	118.49	124.34	129.56	135.82	140.19

Appendix RZR
Table of r to Fisher's z_r

$$z_r = \frac{\log_e\left(\frac{1+r}{1-r}\right)}{2}$$

r	Fisher's z_r	r	Fisher's z_r	r	Fisher's z_r
.00	.000	.34	.354	.68	.829
.01	.010	.35	.365	.69	.848
.02	.020	.36	.377	.70	.867
.03	.030	.37	.388	.71	.887
.04	.040	.38	.400	.72	.908
.05	.050	.39	.412	.73	.929
.06	.060	.40	.424	.74	.950
.07	.070	.41	.436	.75	.973
.08	.080	.42	.448	.76	.996
.09	.090	.43	.460	.77	1.020
.10	.100	.44	.472	.78	1.045
.11	.110	.45	.485	.79	1.071
.12	.121	.46	.497	.80	1.099
.13	.131	.47	.510	.81	1.127
.14	.141	.48	.523	.82	1.157
.15	.151	.49	.536	.83	1.188
.16	.161	.50	.549	.84	1.221
.17	.172	.51	.563	.85	1.256
.18	.182	.52	.576	.86	1.293
.19	.192	.53	.590	.87	1.333
.20	.203	.54	.604	.88	1.376
.21	.213	.55	.618	.89	1.422
.22	.224	.56	.633	.90	1.472
.23	.234	.57	.648	.91	1.528
.24	.245	.58	.662	.92	1.589
.25	.255	.59	.678	.93	1.658
.26	.266	.60	.693	.94	1.738
.27	.277	.61	.709	.95	1.832
.28	.288	.62	.725	.96	1.946
.29	.299	.63	.741	.97	2.092
.30	.310	.64	.758	.98	2.298
.31	.321	.65	.775	.99	2.647
.32	.332	.66	.793		
.33	.343	.67	.811		

Appendix ZRR
Table of Fisher's z_r to r

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Fisher's z_r	r	Fisher's z_r	r	Fisher's z_r	r
.00	.000	.43	.405	.86	.696
.01	.010	.44	.414	.87	.701
.02	.020	.45	.422	.88	.706
.03	.030	.46	.430	.89	.711
.04	.040	.47	.438	.90	.716
.05	.050	.48	.446	.91	.721
.06	.060	.49	.454	.92	.726
.07	.070	.50	.462	.93	.731
.08	.080	.51	.470	.94	.735
.09	.090	.52	.478	.95	.740
.10	.100	.53	.485	.96	.744
.11	.110	.54	.493	.97	.749
.12	.119	.55	.501	.98	.753
.13	.129	.56	.508	.99	.757
.14	.139	.57	.515	1.00	.762
.15	.149	.58	.523	1.01	.766
.16	.159	.59	.530	1.02	.770
.17	.168	.60	.537	1.03	.774
.18	.178	.61	.544	1.04	.778
.19	.188	.62	.551	1.05	.782
.20	.197	.63	.558	1.06	.786
.21	.207	.64	.565	1.07	.789
.22	.217	.65	.572	1.08	.793
.23	.226	.66	.578	1.09	.797
.24	.235	.67	.585	1.10	.800
.25	.245	.68	.592	1.11	.804
.26	.254	.69	.598	1.12	.808
.27	.264	.70	.604	1.13	.811
.28	.273	.71	.611	1.14	.814
.29	.282	.72	.617	1.15	.818
.30	.291	.73	.623	1.16	.821
.31	.300	.74	.629	1.17	.824
.32	.310	.75	.635	1.18	.827
.33	.319	.76	.641	1.19	.831
.34	.327	.77	.647	1.20	.834
.35	.336	.78	.653	1.21	.837
.36	.345	.79	.658	1.22	.840
.37	.354	.80	.664	1.23	.843
.38	.363	.81	.670	1.24	.845
.39	.371	.82	.675	1.25	.848
.40	.380	.83	.680	1.26	.851
.41	.388	.84	.686	1.27	.854
.42	.397	.85	.691	1.28	.856

1.29	.859	1.53	.910	1.77	.944
1.30	.862	1.54	.912	1.78	.945
1.31	.864	1.55	.914	1.79	.946
1.32	.867	1.56	.915	1.80	.947
1.33	.869	1.57	.917	1.81	.948
1.34	.872	1.58	.919	1.82	.949
1.35	.874	1.59	.920	1.83	.950
1.36	.876	1.60	.922	1.84	.951
1.37	.879	1.61	.923	1.85	.952
1.38	.881	1.62	.925	1.86	.953
1.39	.883	1.63	.926	1.87	.954
1.40	.885	1.64	.927	1.88	.954
1.41	.887	1.65	.929	1.89	.955
1.42	.890	1.66	.930	1.90	.956
1.43	.892	1.67	.932	1.91	.957
1.44	.894	1.68	.933	1.92	.958
1.45	.896	1.69	.934	1.93	.959
1.46	.898	1.70	.935	1.94	.960
1.47	.900	1.71	.937	1.95	.960
1.48	.901	1.72	.938	1.96	.961
1.49	.903	1.73	.939	1.97	.962
1.50	.905	1.74	.940	1.98	.963
1.51	.907	1.75	.941	1.99	.963
1.52	.909	1.76	.943		