# Meta-analysis as Judgment Aggregation

## 1. Introduction

The way to approximate total evidence in a wide range of contemporary disciplines, including medical, educational and behavioural sciences, goes by the name of meta-analysis. Typically, medical and social science literature abounds with contradictory results on the same issue: Does vitamin E prevent heart attacks? Does psychotherapy really help people? Do government sponsored job training programs work? Despair over the growing body of conflicting results on such questions was a staple of human sciences until the third quarter of the twentieth century, when the practice of meta-analysis promised to deliver a principled way of resolving conflict among experts. In the passionate assessment of the psychologist Frank Schmidt this new meta-analytic practice shows that

> scientific progress is possible. It means that cumulative understanding and progress in theory development is possible after all. It means that the behavioral and social sciences can attain the status of true sciences; they are not doomed forever to the status of quasi-sciences or pseudosciences. One result of this is that the gloom, cynicism, and nihilism that have enveloped many in the behavioral and social sciences is lifting. Young people starting out in the behavioral and social sciences today can hope for a much brighter future. (Schmidt 1996, 123).

Over the last three decades, the use of meta-analytic tools has grown at a breakneck pace, corresponding to the equally explosive rate of growth in epidemiological research and other experimental studies in human sciences. Integrating evidence by these new quantitative techniques has replaced traditional discursive reviews in social and medical sciences, for it is also claimed that meta-analyses correct for the bias of the reviewer, and help discount misleading evidence as the extant data are merged in a systematic manner.

The first such study is traced back to the work of the psychologist Gene V. Glass who in the 1970s sought to assess the effectiveness of psychotherapy by undertaking a quantitative review of numerous studies carried out to this end. Ever since, Glass's procedure has inspired statisticians—both the theoretical ones and the data-analysts—who have been developing more refined mathematical techniques to integrate evidence. Nowadays equipped with meta-analysis software, the outcome of this activity is what some researchers describe as a meta-analytic revolution. Meta-analyses set new standards of excellence on what counts as strong evidence. While the evidence-based medicine manuals crown this method as the best way of summarizing different research findings, more and more headlines about medical research come from meta-analytic findings.[1] In the current prevailing mood in medical and behavioural sciences, it is only a properly conducted, up-to-date meta-analysis that licenses detachment of hypotheses from the host of evidential claims made in individual studies, which claims may be deemed inconclusive or contradictory with each other.

The meta-analytic procedures achieve this detachment by aggregating available statistical evidence on what the statistician D.B. Rubin calls, in some degree of abstraction, a response surface (Rubin 1992). In Rubin's account, each primary

[1] In the EBM hierarchy of types of evidence, systematic reviews (SRs) of randomized trials is ranked the most trustworthy—"SRs, by combining all relevant randomized trials, further reduce both bias and random error and thus provide the highest level of evidence currently achievable about the effects of health care." (Sacket at all 2000, 134).

study yields a result depending on two sets of situations: the characteristics of the population examined and the design of the experiment (for instance, sample size, sample selection procedure, etc). While each study aims to measure the fixed or variable population characteristics in question by using a nearly ideal experimental design, in reality a study achieves this goal only with errors, accruing from parameter and sampling variations as well as from possible flaws and biases in its experimental design. However, since the goal in each study is the same, namely, finding the true characteristics of the same population, the project of a study is analogous to that of a judge who is asked to pass a judgment on the truth or falsity of a specific charge in the setting of the Condorcet Jury Theorem. Meta-analyses can thus be seen as a type of aggregation of judgments of different research groups working on similar issues, where each group generates its evidence for or against a claim.

My goal in this paper is to see the extent to which judgment aggregation methods subsume meta-analytic ones. To this end, I derive a generalized version of the classical Condorcet Jury Theorem (CJT), the aggregative implications of which have been widely exploited in the area of rational choice theory, but not yet in philosophy of science. According to the theorem the French philosopher Condorcet developed in the hopes of improving the French tribunal system, under some plausible assumptions, the probability that the majority of judges makes a correct decision gets arbitrarily close to unity when the size of the tribunal increases. This theorem acquired new currency in the twentieth century attempts to develop quantitative models of group decision-making. My contention is that the generalized CJT that I prove below is also useful for modelling at least some meta-analytic procedures. I conclude by examining the presuppositions and shortcomings of this model.

Even though the CJT does not necessitate a Bayesian analysis of evidence, I employ the latter in order to evaluate the properties of the inferential framework I present. This gives me the liberty to talk in the same breath both of the acceptance or rejection of a hypothesis--that is, an action space in the manner of classical statisticians--as well as a rational agent's degrees of belief about propositions. I spell out the details of this double use in the course of my exposition below.

## 2. A Condorcet Jury Theorem for Meta-analyses

A meta-analysis is carried out by retrieving and combining the evidence—almost always the final results rather than the basic data—provided by different primary studies, with a view to assessing the overall support for or against a hypothesis in a research area.[2] This procedure is not needed in highly theoretical sciences, such as physics or chemistry, where replications of experiments do not usually exceed a handful, if at al, but meta-analyses are widely used in less theoretical and more empirical sciences, where experimental findings cannot be easily aligned to possible theoretical considerations.[3] Thus, numerous studies have been carried out to examine the relation between cancer risks and diet, the efficacy of teaching methods in

[2] For a popular introduction, see (Hunt 1997) and (Light and Pilllemer 1984). For the statistical methods used, see (Cooper et al 1994), (Hedges and Olkin 1985) and (Hunter and Schmidt 1990). For a criticism of the quantitative spirit it involves, see (Hammersley 2001).

[3] As Nancy Cartwright observes, "In physics there is a rich network of knowledge and a great deal of connectedness so that any one hypothesis will have a large number of different consequences by different routes to which it is answerable. This is generally not true of hypotheses in the social sciences. In social sciences, we need techniques to export conclusions from where they are confirmed to across the board". (Cartwright 2007, 74). Cartwright holds that in social sciences there is no rigorous justification for exporting results from the populations and situations in which they are established.

mathematics education, etc. These studies need not be replications of each other—they usually sample from different populations, use different research designs, and moreover may address different sets of questions. In meta-analyses, at least one question or issue is singled out and the cumulative evidence concerning that issue is assessed, pooling results from studies that satisfy some explicitly stated criteria of inclusion concerning experimental design.

A simple model for this kind of post-data analysis is the following: Let *H* stand for the hypothesis that is tested, and suppose there are *n* studies, each of which provides a binary evaluation of *H*, for instance, accepting *H* or rejecting *H*. We can codify each study outcome as an indicator function $S_i$ which takes the value 1 if the $i^{th}$ study accepts *H* or 0 if the $i^{th}$ study rejects *H*. As it is customary in statistical tests with a pre-data evaluation scheme, the probabilities with which these values are realized *conditional* on whether *H* is or is not the case can be specified as an error-statistical property of the primary testing procedure—the first corresponds to the reliability (=1-size) and the second to the power of the test. Suppose that there is a good estimate of the reliability and the power of each study, so that for each *i*=1, ..., *n*, the following are well-defined probabilities:

$$P(S_i=1| H)=r_i.$$

$$P(S_i=0| \neg H)=s_i.$$

Even though the above assumptions derive from the framework of standard classical tests, in which the probability measure *P* is regarded objective, I believe there is no reason not to consider *P* as measuring the degrees of beliefs of an agent appraising the test design for the hypothesis *H*.

How can we merge the information from *n* such binary test results which may possibly—and in practice typically—yield incongruous results? Here is an analysis inspired by the classical CJT: Let $S(n)=S_1+\ldots+S_n$, so that *S(n)* counts the number of studies which accept *H*. Note that *S(n)* is a random variable, the conditional distribution of which *given H* (or *given* $\neg$ *H*) can be expressed in terms of $r_i$'s (or $s_i$'s). Assume that the primary studies are independent from one another, conditional on the truth value of H.[4] Even though the independence condition can be relaxed in favor of a limited amount of dependence between the primary tests, I will not consider this more general case here partly for lack of space but mostly for the sake of highlighting the generalization to CJT that I propose.[5] If we further assume

(1) uniform reliability, that is, $r_1=\ldots=r_n=r$

then *S(n)* has the binomial distribution $b(r, n)$ conditional on *H*. Similarly, if we assume

(2) uniform sensitivity, that is, $s_1=\ldots=s_n=s$

then *S(n)* has the binomial distribution $b(1-s, n)$ conditional on $\neg H$. By the (weak) law of large numbers, if we assume (1), then *S(n)/n* approaches in probability to *r*, given *H*. Similarly, if (2) is assumed, then *S(n)/n* approaches in probability to *1-s*, given $\neg H$. In terms of formulae this means:

a) If (1), given any $\varepsilon > 0$, $P(\left|\frac{S(n)}{n} - r\right| < \varepsilon \mid H)$ converges to 1 as *n* tends to infinity.

Similarly,

[4] I discuss the implications of this condition further below.

[5] To see generalizations of the CJT in this direction, see (Hawthorne 1991) and (Resnick 1998, 270-74).

b) if (2), given any ε > 0, $P(\left|\frac{S(n)}{n} - (1-s)\right| < \varepsilon \mid \neg H)$ converges to 1 as *n* tends to infinity.

Suppose that after surveying *n* primary studies, one decides to accept *H* if and only if the frequency of acceptances exceeds a fixed ratio *c*. This decision rule induces a new random variable $A_n$, which I will refer to as the c-aggregation rule, with the following indicator function:

$$A_n = \begin{cases} 1 & \text{if } \frac{\mathrm{S(n)}}{\mathrm{n}} > c \\ 0 & \text{otherwise} \end{cases}$$

There may be several other aggregation rules, for instance those based on absolute thresholds, but for simplicity I examine only this one in this paper. Due to the conditional asymptotic behavior of *S(n)/n* mentioned above, it can be shown easily that we have the following results:

i) If (1) holds and c < r, then $P(A_n=1 \mid H)$ converges to 1 as *n* tends to infinity.

ii) If (1) holds and c > r, then $P(A_n=1 \mid H)$ converges to 0 as *n* tends to infinity.

iii) If (2) holds and c < 1-s, then $P(A_n=0 \mid \neg H)$ converges to 0 as *n* tends to infinity.

iv) If (2) holds and c > 1-s, then $P(A_n=0 \mid \neg H)$ converges to 1 as *n* tends to infinity.[6]

[6] See appendix for the proof.

So a necessary and sufficient condition for $A_n$ to yield asymptotic consistency with the true state of the world, that is, for it to be unbiased, is that (1) and (2) hold and the cut-off value *c* satisfies:

$$1\text{-s} < c < r$$ [7]

A necessary and sufficient condition for the last inequalities is that $1\text{-s} < r$, or that $1 < s + r$, given that the cut-off value *c* can be adjusted subsequently to the determination of error probabilities. The classical Condorcet Jury Theorem is a special case when $r=s>1/2$, and $c=1/2$.[8] The literature on the extensions of the CJT, as far as I know, accepts the condition $r=s>1/2$ unquestioningly, perhaps because in the rational choice theory applications it is customary to assume competent decision makers in this minimal sense. In the meta-analysis application that I envision, we do not need to assume such 'competent' primary tests. A primary study can be considered good enough if for that study the related reliability and sensitivity satisfy: $1 < s + r$. Thus a primary study with reliability 0.9 and sensitivity 0.2 can be pooled for the purposes of this kind of meta-analysis, if one uses, for instance, an 0.85-aggregation rule.

One can similarly investigate the asymptotic behavior of $P(H| A_n=1)$ and $P(\neg H| A_n=0)$ as *n* tends to infinity. These probabilities are crucial in a Bayesian analysis of the c-aggregation rule $A_n$, to assess whether an agent using this rule can succeed in evaluating the cumulative evidence for *H* correctly when *n* is sufficiently large. Here I switch to the viewpoint of a rational agent whose degrees of belief

[7] This follows from the observation that $A_n$ is unbiased if and only if in the long run it indicates with nearly unit probability that *H* is (not) the case if *H* is (not) the case. The latter is equivalent to requiring that the results i) and iv) hold, while the results ii) and iii) do not.

[8] For a historical exposition of the CJT, see (Daston 1988, 340-352).

concur with the error probabilities of primary tests.[9] Assuming that this agent holds *H* with a prior probability *x* strictly between 0 and 1, we have by Bayes's theorem:

$$P(H \mid A_n = 1) = \frac{xP(A_n = 1 \mid H)}{xP(A_n = 1 \mid H) + (1-x)P(A_n = 1 \mid \neg H)}$$
$$= \frac{1}{1 + \left(\frac{1-x}{x}\right)\frac{P(A_n = 1 \mid \neg H)}{P(A_n = 1 \mid H)}}$$

There are many sets of sufficient conditions for this quotient to converge to 1. For instance, we can stipulate the following to ensure this:

> $0 < x < 1$; uniform sensitivity (i.e. condition (2)); $c > 1\text{-}s$; a positive lower bound on the posteriors of $A_n=1$, (for instance, $\liminf_n P(A_n=1| H) \neq 0$).[10]

An analogous analysis shows that $P(\neg H \mid A_n=0)$ converges to 1 provided that $0< x <1$; condition (1); $c < r$ as well as something to the effect that $\liminf_n P(A_n=0| \neg H) \neq 0$.

Conjoining these sets of sufficient conditions, it follows that the rational agent described above would be successful in the long run with her inferences on the basis of the c-aggregation rule, in the sense that this rule indicates her the truth or falsity of *H* with a high degree of probability, provided that the conditions $0 < x < 1$; (1); (2) and $1\text{-}s < c < r$ hold. (Note that when we combine the two inequalities, it is guaranteed that $\liminf_n P(A_n=1| H) \neq 0$ and $\liminf_n P(A_n=0| \neg H]) \neq 0$).

[9] That a rational agent should do so can be argued on the basis of Lewis's Principal Principle. Of course, not being a theorem of the probability theory, the Principal Principle is an additional assumption available to those who seek to establish a plausible connection between degrees of belief and chances.

[10] These conditions jointly entail the result iv) above, as well as ensuring that the ratio in question is well-defined in the limit.

One can further improve these findings by employing another generalized version of the CJT, where the uniformity of reliabilities (1) or of sensitivities (2) are weakened by the following conditions:

(1)' $\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{i=n} r_i = \bar{r}$

(2)' $\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{i=n} s_i = \bar{s}$

That is, instead of requiring uniformity of reliabilities and sensitivities of the studies, we require their convergence. This is a more general condition that includes the former ones (i.e., (1) implies (1)').[11] In this setting, both consistency and the success of the c-aggregation rule in the long run are ensured on the condition that:

$1 - \bar{s} < c < \bar{r}$

Since the c-aggregation rule can be fixed after having some idea of the asymptotic behavior of $s_i$'s and $r_i$'s, the crucial condition here is again the inequality

$1 < \bar{r} + \bar{s}$.

## 3. Presuppositions of the Model

The above model shows that instead of the sophisticated meta-analyses encountered in the literature, one can simply do a vote-counting analysis of primary studies if certain conditions are satisfied. These are:

a) Publication bias or other kinds of bias do not exist.

[11] For a proof, see (Resnick 1998, 205).

b) There are sufficiently many primary studies that are independent of each other (given the true state of the world).

c) $1 <$ limiting reliability + limiting sensitivity.

The first assumption is important not only for the application of this model to meta-analyses, but *for any other meta-analytic practice as well*. Publication bias refers to the dependence of the publication of a research on its result. Many scientific journals resist the publication of non-significant results (in the context of classical statistical tests). If there is such a bias, then many, if not all, published results do not transmit information to the effect that H. In other words, most of the studies which would signal S=1 are eliminated from the available pool of studies—in the meta-analytic jargon this corresponds to the fugitive literature, most of which is (or used to be) kept in file-drawers. If there is a publication bias, then

$$P(S=1|H) \neq P(S=1||H \text{ and } S \text{ is published}),$$

and hence the above model cannot be applied to meta-analyses that retrieve only published results. A related problem for the meta-analyst is accessing studies that are not published for other reasons than the publication bias, for instance, dissertations or government studies. If such studies exhibit other sorts of tendencies in their results, then the meta-analyses which exclude them could not aspire to represent the total evidence. Because the direction of bias can be accurately quantified from publication policy practices, statisticians have devised ways of detecting publication bias and corrections thereof (See, for instance, (Duval and Tweedie 2000). My model, as it stands, cannot take into account corrections for these kinds of biases.

Concerning the assumption on the number of studies, it is certainly unrealistic in many disciplines to assume there are even thousands of them. Yet, the convergence rates of binomial variables are fairly rapid, and we can usually determine an upper bound on the number of studies needed to ensure desired levels of approximations to the limiting values. Furthermore, in the current state of the art in empirical research, it would also be incorrect to underestimate the amount of quantitative primary research. This is the day is of the empirics in the human sciences: Scientific journals are awash in data; hospitals, schools and government offices are pouring figures and numbers, not to mention biotech and financial companies or myriads of other databases. To give an example concerning the topic of the earliest meta-analysis of Smith and Glass in 1977, based on 375 studies, there have been so many follow-ups that some researchers had to meta-analyze the extant meta-analyses! Already in 1993 a vast study analyzed 302 meta-analyses of a total of nearly five thousand primary studies, followed by numerous other studies on the same. The figures are ever on the rise.[12]

The second assumption, namely that the primary studies are independent from each other given the true state of the world, is usually the case. Researchers may interact with each other but so long as they generate their own evidence and assessments through their research—which activity can be taken as definitive of primary research—we can treat their findings as independent verdicts on $H$.[13] The fact that each primary research generates its own evidence is a feature that distinguishes this application of the generalized CJT from its standard applications in the rational choice theory. In the original setting of the CJT, each judge evaluates the same body of evidence for the tribunal case at issue. This is not so in the meta-analytic practice, where each primary research creates a separate body of evidence

[12] See (Hunt 1997, 43).

[13] See (Hawthorne 2001) for a similar point about CJT.

regarding the hypothesis $H$. As my model indicates, however, there is no need to assume that the same body of evidence is available to each primary researcher. The independence assumption prevents duplications of a single study (for instance, reported in different sources) being treated as different pieces of research, but it does not block replications of the same experiment (carried on different samples or even on the same sample but with different experimental design) to count as different pieces of research. Standard meta-analyses also include the latter.

What if each primary research is systematically biased because of some flawed background assumptions or measurement procedures? History of science is full of episodes where many conceptual, interpretative or practical commitments of scientific communities prevented them from seeing through the veil of paradigms. One should then not expect to infer from unanimous agreement on $H$ (i.e., when $S(n)=n$) to $H$ with full assurance. There are several ways in which the CJT type model above can be refined to capture this situation. The simplest way would be to incorporate all such shared background assumptions as a statement $B$ on which the agents conditionalize their beliefs. In other words, instead of using the probability measure $P$, one can use the probability measure $P(\ |B)$ or in more compact notation, $P_B$. If the reliabilities and the sensitivities measured with $P_B$ still satisfy the crucial conditions stated in the above model, then the same conclusions would ensue. The heart of the problem is whether one can indeed quantify reliabilities and sensitivities with the same accuracy when $B$ is assumed to distort assessments of how raw data or evidence bears on hypotheses.[14] *I believe this is a problem, not only for my model but also for any meta-analysis*.

The third assumption points to another problem that may arise in the test designs of primary studies. Each pair of numbers $r_i$ and $s_i$ have to do with the error characteristics of the $i^{th}$ study. In the design of classical statistical tests, one of these numbers is first fixed at a desirable level and then the other is optimized later. In other words, one usually cannot design a test in which both of these values are preset at any two desirable levels. Hence, usually one and only one of the assumptions (1)' or (2)' can be realized in practice. Precisely because of this, it is not a priori certain that $1 < \bar{r} + \bar{s}$, even though one may grant that for many tests $1 - s_i < r_i$, as the former quantity stands for Type II error and the latter for the reliability (=1-Type I error) of a test. Notoriously, many significance tests have very low power, and the above inequality would not hold for such tests. On the other hand, we may take the condition $1 - s_i < r_i$ as one of the inclusion criteria of a meta-analysis. In other words, if the condition $1 - s_i < r_i$ is not satisfied for a primary study, that study may be excluded from the meta-analytic review, a requirement that statisticians would probably concur with.

[14] Another way to solve this problem is suggested in (Dietrich and List 2004). Instead of modeling the meta-analytic procedure as an evaluation of all sorts of possible evidence the world presents to the researchers, we can model it as an evaluation of the body of possible evidence the research findings exhibit with constraints deriving from paradigmatic commitments. The difference amounts to determining whether the world was the immediate cause of the results $S_i$'s or whether the evidential framework *E* preempts the world's input. If the latter is the case, then it can be inserted as a new random variable *E* between the state of the world {H, ¬H} and the study outcomes $S_i$'s in such a way that *E* screens off the former in $S_i$'s ancestry. In this case, the convergence results mentioned in my model should be modified so that they reveal in the limit the misleading role of *E*.

## 4. Conclusion

The model I propose is too simple, and thereby has both advantages and disadvantages. To start with, the vote-counting method is not too selective about experimental design. This can be seen as an advantage when the dispute over the nature of best experimental designs--for instance, between Bayesians and classical statisticians—is taken into account. Yet, this is also a disadvantage. The most serious shortcoming of this model is that it assumes all primary studies are on a par so long as their error probabilities satisfy the condition $1<r+s$. Yet, a proper meta-analysis begins by classifying the included primary studies with a view to rank them on the basis of their experimental design. In the current practice of meta-analysis in many disciplines, RCTs with larger samples are given more weight than small-scale studies, or non-randomized studies. The scientific community would rightly hesitate between accepting the result of a meta-analysis and a conflicting result from a new large-scale study.[15] They would not treat the latter study as having merely one *n*th of a vote. I do not suggest that vote-counting always gets us closer to truth, but there may be ways to address this problem. One way is to follow the procedure of a proper meta-analysis by first stratifying the extant research using some epistemic values, and then applying my model to each stratum separately. That will work provided that there are sufficiently many primary studies in each category of the hierarchy. Another shortcoming of my model is that it cannot be used to determine effect size, the quantity most current primary research is driven to reveal. Another issue about which my model is silent is the discovery of new evidence through the meta-analytic

[15] This situation is quite frequently the case in clinical research. As (Ioannidis 2006) counts, 16% of highly cited primary research was plainly contradicted by subsequent research, and another 16% was found to have exaggerated the effects of medical interventions.

practice, when statisticians are keen to determine the reasons for conflicting results by conjecturing and testing for moderator variables. No doubt, any binary evaluation of the results of primary studies is doomed to dissipate valuable information deriving from empirical research.[16] On the other hand, we have to find ways to extract information from the ever-increasing profusion of empirical research. We live in an era of massive efforts to collect data, on scales unimaginable until one century or so, and it may be expedient to extract reliable evidence from data in simple and manageable ways. What I have in mind is in the spirit of explorative data analysis envisioned by the statistician John W. Tukey. Sometimes stem-and-leaf plots are more informative than sophisticated mathematical tools.[17] I believe that sometimes, and at least under the conditions specified in the above model, simple judgment aggregation procedures can subsume meta-analysis.

## Appendix

In order to prove i), note that when c < r, we have the following:

$$
\begin{aligned}
P(A_n = 1 \mid H) &= P(\frac{S(n)}{n} > c \mid H) \\
&= P(r - \frac{S(n)}{n} < r - c \mid H) \\
&\geq P(c - r < r - \frac{S(n)}{n} < r - c \mid H) \\
&= P(\left|\frac{S(n)}{n} - r\right| < r - c \mid H)
\end{aligned}
$$

[16] Cooper (1994, 24) distinguishes between "review-generated evidence," and the "study-generated evidence" to highlight the significance of the former. Review-generated evidence may include the gender of the researchers and the publication date (of primary studies).

[17] Tukey's project was precisely to de-emphasize formal mathematics in data analysis. See (Donoho 2000) for an assessment of Tukey and the future of data analysis.

When in addition (1) holds, as mentioned in the text, the last expression converges to 1 when *n* tends to infinity. This establishes the same for P($A_n$=1|H).

To prove ii), assume c > r, and note the following:

$$
\begin{aligned}
P(A_n = 1 \mid H) &= P(\frac{S(n)}{n} > c \mid H) \\
&= P(\frac{S(n)}{n} - r > c - r \mid H) \\
&= 1 - P(\frac{S(n)}{n} - r \leq c - r \mid H) \\
&\leq 1 - P(\left|\frac{S(n)}{n} - r\right| \leq c - r \mid H) \\
&\leq 1 - P(\left|\frac{S(n)}{n} - r\right| < c - r \mid H)
\end{aligned}
$$

If (1) holds, then the last quantity converges to 0, and hence so does P($A_n$=1|H). The remaining results iii) and iv) can be proved in an analogous fashion by replacing *c* with *1-s*, and *H* with its negation.

# REFERENCES

Cartwright, Nancy. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Cooper, Harris and L.V. Hedges eds. 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Daston, Lorraine. 1988. *Classical Probability in the Enlightenment*. Princeton: Princeton University Press.

Dietrich, Franz, Christian List. 2004. A Model of Jury Decisions where All Jurors have the Same Evidence. *Synthese* 142(2): 175-202.

Donoho, David L. 2000. "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality". Lecture delivered at the American Mathematical Society Conference, "Mathematical Challenges of the Twenty-first Century", available online at http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html

Duval, S., & Tweedie, R. 2000. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

Hammersley, Martyn. 2001. On 'Systematic' Reviews of Research Literatures: a 'narrative' response to Evans&Benefield. *British Educational Research Journal* 27*(5)*:543-554.

Hawthorne, James, unpublished manuscript (circulated beginning 2001), "Voting in Search of the Public Good: the Probabilistic Logic of Majority Judgments", online at http://faculty-staff.ou.edu/H/James.A.Hawthorne-1/Hawthorne--Jury-Theorems.pdf

Hedges, Larry V., Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Academic Press: San Diego.

Hunt, Morton. 1997. *How Science Takes Stock: The Story of Meta-Analysis*. Russell Sage Foundation: New York.

Hunter, J. E., F.L. Schmidt. 1990. *Methods of meta-analysis*. Newbury Park: Sage.

Ioannidis, John P.A. 2005. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA,* 294 (2):218-228.

Light, Richard J. and David B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, Massachusetts, and London: Harvard University Press.

Resnick, Sidney I. 1998. *A Probability Path*. Basel: Birhauser Verlag AG.

Rubin, Donald B. 1992. Meta-Analysis: Literature Synthesis or Effect-Size Surface Estimation? *Journal of Educational and Behavioral Statistics* 17: 363-374.

Sackett, David L. et al. 2000. *Evidence-Based Medicine*. Edinburgh: Churchill Livingstone.

Schmidt, Frank L. 1996. Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Pscyhological Methods*, 1(2): 115-129.

Smith, Mary L.; Glass, Gene V. 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist*. 32(9): 752-760.