Journal of Official Statistics, Vol. 15, No. 4, 1999, pp. 477–494

Bayesian Estimation of the Number of Unseen Studies in a Meta-Analysis

Lynn E. Eberly[1] *and George Casella*[2]

Public policies based on science today are often formed with the help of a meta-analysis, a combining of those experiments testing the hypothesis of interest. For example, a nation's drug regulatory body may combine clinical trial results from public health clinics, university research clinics, and pharmaceutical companies to determine if a drug under study is efficacious and safe. However, a parameter estimate from a meta-analysis is biased when the experiments to be combined are a non-random sample from the population of all experiments done on the hypothesis of interest. In particular, publication bias occurs when studies with significant results are more likely to be published than studies with non-significant results. We develop a model for the distribution of the total number of studies carried out, both published and unpublished, dependent on the probability of publication. We assume a selection model where all studies significant at level α are published, while non-significant studies are published with probability ρ. Using a Bayesian hierarchical model with Metropolis simulation and Gibbs sampling techniques, we study how the distribution of the total number of studies changes as ρ changes. An application on lead exposure and IQ level in children is presented and the results interpreted. Comparisons are made with Rosenthal's fail-safe N estimators and with the recent frequentist estimation method of Gleser and Olkin.

Key words: Hierarchical models; file-drawer problem; Gibbs sampling; publication bias; Rosenthal's fail-safe number.

1. Introduction

Meta-analysis, a method of combining results from different experiments testing the same hypothesis, has gained wide recognition in both the statistical and the scientific worlds in the past twenty years. As stated by the United States National Research Council, ''Combining information from disparate sources is a fundamental activity in both scientific research and policy decision making'' (1992, p. 5). For example, the United States National Institute of Health has a Consensus Development Program which produces consensus statements on important topics in medicine. The use of meta-analysis is integrated into the

[1] Division of Biostatistics, University of Minnesota, 420 Delaware Street SE, Box 303, Minneapolis, MN 55455-0378, U.S.A. E-mail: Lynn@biostat.umn.edu
[2] Department of Biometrics, Cornell University, 435A Warren Hall, Ithaca, NY 14853, U.S.A. E-mail: ge15@cornell.edu
Acknowledgments: This article refers to Division of Biostatistics Research Report # 97-018. Biometrics Unit Manuscript # BUM-1308.
The work by Lynn E. Eberly was supported in part by National Institute of Environmental Health Sciences Training Grant EHS-5-T32-ES07261-03 and National Science Foundation Grant DMS-9305547.
The work by George Casella was supported in part by National Science Foundation Grant DMS-9305547.

consensus development process. Recent statements have focused on breast cancer screening, acupuncture, management of Hepatitis C, and many others (http://odp.od.nih.gov/consensus).

The goal of the meta-analysis may be to test a combined effect estimate, meaning attention must be paid to the validity and reliability of that estimate. The most common method of finding experimental results to include in a meta-analysis is through literature searches in relevant journals and dissertation abstracts. However, journals can be unrepresentative for a number of reasons. Often studies with statistically nonsignificant results are underrepresented in the literature. For example, a scientist may not submit the results of a study that does not show some statistically significant result, or a journal editor may not accept those results, either one feeling that a result of ''no difference'' would be of little importance to the scientific community. Thus, any sample of studies from the published literature is typically nonrandom. When a meta-analysis of these studies is then done, an overall effect estimate could be biased towards a higher level of significance (Hedges 1992). According to Bayarri and DeGroot (1986), *selection bias* is the distortion in an effect estimate resulting when a nonrandom sample is drawn from the population of interest. This article focuses on *publication* bias in particular, the selection bias resulting when studies statistically significant at some level α are more likely to be published than nonsignificant studies.

Easterbrook, Berlin, Gopalan, and Matthews (1991) carried out a retrospective study of 285 analyzed research projects which had been approved by the Central Oxford Research Ethics Committee between 1984 and 1987 in order to show that publication bias does in fact exist in the medical literature. Using logistic regression and adjusting for relevant covariates, they found that projects with statistically significant results (defined to have a p-value < 0.05) were more likely to have been published and/or presented than those with nonsignificant results (odds ratio $= 3.56$, 95% C.I. $= (1.82; 6.99)$). In addition, they noted that 43 of the 78 unpublished projects had obtained null results. Only eight of those 43 were written up and subsequently rejected, while 26 were never written up *because* they showed null results. Dickersin, Min, and Meinert (1992) carried out a similar study using research projects that appeared on the institutional review board logs for the Johns Hopkins Health Institutions. Using logistic regression and adjusting for covariates, they found similar results, including the conclusion that the problem lies with authors, not editors.

A variety of methods for dealing with publication bias have been proposed. Rosenthal (1979) began with the fail-safe number, which calculates the number of unseen studies averaging null results needed to bring a meta-analytic result to some pre-specified level of significance. White (1982) and Glass, McGaw, and Smith (1981) suggest obtaining results for studies which were not published (through surveys of colleagues, for example, or national registries of studies) and comparing those results to the published results. Light and Pillemer (1984) describe a method to detect publication bias using a ''funnel graph'' of sample size versus effect estimate. In the presence of publication bias, and assuming effect size is unrelated to sample size, the graph should be missing the lower left-hand corner of the pyramid. Berlin, Begg, and Louis (1989) introduce a method to quantify the information in a funnel graph by using a model relating bias to sample size under the same assumption. Results indicated that small trials are more prone to publication bias and that the bias may be substantial, especially when the trial was based on

a nonrandomized design. A more recent extension of the funnel graph idea (in Begg and Mazumdar 1994) suggests calculating and then testing a rank correlation between effect estimates and their variances. A positive correlation would indicate that negative studies are less likely to be published.

The funnel graph and correlation approaches have the advantage of being based on assumptions which are distribution-free. Hedges (1984) meanwhile pursued truncated sampling models, where it was assumed that statistically nonsignificant results (at α-level $= 0.05$) do not get published. He found that the bias can depend on a study's sample size and effect size, and can be substantial for either small samples or small effects. Bayarri and DeGroot (1986, 1991) explore the behavior of published results using an indicator function of statistical significance to weight the model's likelihood, and show that significant overall results obtained from published data actually can be strongly supportive of the null hypothesis. Iyengar and Greenhouse (1988) modify Bayarri and DeGroot's method slightly by not restricting the selection to this ''publish if and only if significant'' situation. They incorporate a family of weight functions into the model's likelihood, using the conditional probability of reporting a study given the data as the weight, where this probability varies across studies. Hedges (1992) and Dear and Begg (1992) take the same approach, but modify these weight functions slightly, while Cleary (1996) computes estimates of the parameter of interest as a function of the selection parameter. Frongillo (1991) and Silliman (1997) take a Bayesian approach and use two-stage hierarchical models to model variability both within and between studies. Gleser and Olkin (1996) revisit Rosenthal's attempts to explore the number of unpublished studies and introduce several frequentist methods for interval estimates thereof. As the authors point out, these methods take advantage of the fact that under the null hypothesis of interest, p-values from experiments testing this H_0 have a common known distribution which is independent of each experiment's design, sample size, and concomitant variables.

Historically, then, there have been three general methods of dealing with publication bias: truncated sampling models, invariant sampling, and source augmentation. Truncated sampling models assume that no nonsignificant studies are published, and then, usually through simulations, determine the bias in the effect estimate that comes about due to the publication process. Recently this has been extended to include less strict selection processes. Invariant sampling methods limit the meta-analysis to that subset of studies which come from a sampling frame independent of the publication process (e.g., registries of studies); extensive registries of studies, though, do not as yet exist in most fields of research. Source augmentation speculates on the number of missing (unpublished) studies and may then adjust effect estimates accordingly (Begg and Berlin 1988). Of the three methods, truncated sampling and invariant sampling often assume that the researcher has access to each study's effect estimates and perhaps sample variances. Reality forces us to acknowledge, though, that often we cannot acquire the original data from a study, sometimes not even the effect size estimates. Especially with older studies, it is likely that only p-values or test statistics such as t-values can be gleaned from the publication itself; this renders the use of many of the above methods impossible. On the contrary, the source augmentation methods that have been developed so far (as well as the one we will explore) do not require more than p- or t-values. In spite of this advantage, we

believe source augmentation should, whenever possible, be carried out *in addition to* effect size estimation. Both are important aspects of a meta-analysis.

In this article, we use a hierarchical Bayesian structure to model the distribution of the total number of studies carried out, both seen and unseen, dependent on the probability of publication. This method still necessitates estimating a selection probability, but the distribution can then be calculated for a range of probability values, leading at least to a somewhat more detailed picture.

Section 2 of this article covers the derivation of the model and the assumptions associated with it, including the sampling methods used. Section 3 explains the results from simulations based on the model. Section 4 presents an application of the results to a meta-analysis on studies of lead exposure and IQ levels in children, and makes comparisons to Rosenthal's and Gleser and Olkin's source augmentation methods. Section 5 presents our conclusions regarding the uses and limitations of this theory, and directions for further research.

2. The Approach

2.1. The model

Throughout this article, we assume that some of the assumptions necessary to conduct a meta-analysis hold: (i) each of the observed studies tests the same hypothesis; (ii) the observed studies are independent. The following is a usual assumption of meta-analysis that we presume does *not* hold: (iii) the observed studies are a random sample from the population of all studies that have been carried out on this hypothesis. Most researchers agree that *some* form of selection bias, particularly publication bias, is present in any field, which invalidates assumption (iii). The probability of publication, call it Q, quite likely varies widely from field to field, from journal to journal, and maybe even from year to year. We will impose a prior Beta distribution on Q in order to account for this variability:

$$\pi_Q(q|a,b) = q^{a-1}(1-q)^{b-1}/B(a,b) \quad 0 \le q \le 1, a, b > 0$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(x) = \int_0^\infty t^{x-1}\exp^{-t} dt$. (Throughout this article, we will use the symbol π to denote probability mass or density functions.) The Beta distribution is very flexible, and by its parameters can vary from a bathtub shape through a uniform to a bell-shaped distribution. Assuming publication bias is present, Q must be dependent on the probability of achieving a statistically significant result, call it R. By the laws of probability, we can write:

$$\begin{aligned} Q &= P(\text{publication}|\text{significant})\,P(\text{significant}) \\ &\quad + P(\text{publication}|\text{nonsignificant})\,P(\text{nonsignificant}) \\ &= R + \rho(1-R) \end{aligned} \tag{1}$$

This structure dictates that all significant studies are published, and that some proportion $0 \le \rho \le 1$ of the nonsignificant studies are published. ρ is a selection parameter; we will treat it as a known value.

When conducting a meta-analysis, one reviews the available literature and finds all published studies that test the hypothesis of interest. If k such studies are found, there are still

an unknown number, call it $N - k$, of studies that have actually been done, but were *not* published. We can thus model N using a Negative Binomial distribution: how many studies does it take until we see k successes with success probability $Q = q$?

$$\pi_N(n|q,k) = \binom{n-1}{k-1} q^k (1-q)^{n-k} \quad n = k, k+1, \ldots, \quad 0 \le q \le 1. \tag{2}$$

The distribution of N that we have here is conditional on an unknown value, namely q. What we are ultimately interested in is that marginal distribution of N which is no longer dependent on the value of q.

It is easy to find this marginal through the calculation $\pi_N(n|k) = \int \pi_N(n|q,k)\pi_Q(q)dq$, but the only observed data that this incorporates is the number of published studies, k. We are ignoring important information relevant to publication bias if we do not take into account the number of *significant* published studies and the structure of Equation 1. As we shall see in Section 2.2, incorporating this information makes it much less straightforward to find $\pi_N(n|k)$. We will obtain $\pi_N(n|k)$ through a Gibbs sampling procedure (as will be explained in Section 2.3), but this procedure requires our model's *full conditional specification* (FCS):

$$\pi_N(n|q,\theta,\text{data}) \quad \text{and} \quad \pi_Q(q|n,\theta,\text{data}) \tag{3}$$

These are the conditional distributions of each unknown parameter of interest, where θ denotes the nuisance parameters (ρ, a, b) and ''data'' denotes the number of observed studies and the number of those which show significant results at level α.

2.2. *Derivation of the full conditional specification*

We begin with the more complicated piece of the FCS, the conditional distribution of Q given $N = n$. Using $\pi_N(n|q,k)$, $\pi_Q(q|a,b)$, and Bayes's rule, we can show that:

$$\pi_Q(q|n,k,a,b) = q^{k+a-1}(1-q)^{n-k+b-1}/B(k+a, n-k+b)$$

We still need to incorporate the observed number of significant studies. Consider the formulation of Q given in Equation 1. Given $R = r$ and a pre-specified level of significance α, any study will be significant with probability r and nonsignificant with probability $1 - r$. Note that $r = \alpha$ only when H_0 is truly correct; otherwise $r > \alpha$. Assuming studies are independent (which is not too unreasonable), every study done is the realization of a Bernoulli(r) random variable. (Since larger studies will have more power, and hence are actually more likely to achieve statistical significance, we need to assume that the studies are of approximately the same size; then r will be constant across studies. This issue will be discussed more in Section 5.) The k observed studies in particular are thus k independent Bernoulli trials, of which a certain number will be ''successes,'' where a success means statistical significance. This leads us to a Binomial(k, r) random variable, call it Z, which counts the number of *significant* studies within the *observed* studies:

$$\pi_Z(z|k,r) = \binom{k}{z} r^z (1-r)^{k-z} \quad z = 0, 1, \ldots, k, 0 \le r \le 1$$

The usual estimate of a Binomial probability is $\hat{r} = z/k$, the maximum likelihood estimator.

We can then get estimates of Q using Equation 1: $\hat{Q} = \hat{r} + \rho(1 - \hat{r}) = (1 - \rho)z/k + \rho$. We can now calculate the distribution of $\hat{Q}$, dependent on the values of q, k, r, and ρ:

$$\begin{aligned}\pi_{\hat{Q}}(\hat{q}|q,k,r,\rho) &= P[(1-\rho)Z/k + \rho = \hat{q}|k,r,\rho] \\ &= P\left[Z = \frac{k(\hat{q}-\rho)}{1-\rho}\middle| k,r,\rho\right] \\ &= \binom{k}{\frac{k(\hat{q}-\rho)}{1-\rho}} r^{\frac{k(\hat{q}-\rho)}{1-\rho}}(1-r)^{k-\frac{k(\hat{q}-\rho)}{1-\rho}}\end{aligned}$$

where $\rho \leq \hat{q} \leq 1$ and $0 \leq \rho \leq 1$. Although it is not explicitly part of the equation, this density *is* dependent on q, through r and ρ by Equation 1. Note also that the introduction of the dependence on ρ leads the range of $\hat{q}$ to be bounded below by ρ.

Now that we have derived the distribution of the observed data $\hat{q}$, we should incorporate that information into our full conditional specification (Equation 3). Again using probability calculus:

$$\pi_Q(q|n,k,\rho,a,b,\hat{q}) = \frac{(q-\rho)^z(1-q)^{n+b-z-1}(q)^{k+a-1}}{\int_\rho^1 (q-\rho)^z(1-q)^{n+b-z-1}(q)^{k+a-1}dq} \qquad \rho \leq q \leq 1 \tag{4}$$

We now have $\pi_Q(q|n,\theta,\text{data})$ and $\pi_N(n|q,k)$. It appears as if the conditional distribution that we have for N is not the distribution needed for the full conditional specification, but note that given a value for q, the values of ρ, a, b, and $\hat{q}$ are irrelevant. In other words, we assume that N is conditionally independent of these values. Thus, $\pi_N(n|q,\theta,\text{data}) = \pi_N(n|q,k)$ and we are ready to implement the Gibbs sampler. To simplify notation, and since they are assumed known, we will suppress the dependence on ρ, a, and b from now on.

2.3 Sampling techniques

The Gibbs sampler is an iterative Markov chain Monte Carlo simulation technique introduced by Geman and Geman (1984) and further developed by Tanner and Wong (1987) and Gelfand and Smith (1990). A gentle introduction can be found in Casella and George (1992). Very generally speaking, the purpose of the Gibbs sampler is to replace a difficult calculation (here, of $\pi_N(n|k)$) with a sequence of easier calculations (using $\pi_N(n|q,k)$). The algorithm alternately generates values from our two distributions in Equation 3 as follows:

[0.] Choose an arbitrary starting value $q_o \in [0,1]$.

[1.] For $i = 1,\ldots,t$, generate : n_i from $\pi_N(n|q_{i-1},k)$

q_i from $\pi_Q(q|n_i,k,\hat{q})$

Under regularity conditions described in Geman and Geman (1984) and Tanner and Wong (1987), among many others, the values of n_i and the values of q_i over the iterations form two Markov chains, $n_1, n_2, \ldots, n_t$ and $q_1, q_2, \ldots, q_t$. We then also have the following

asymptotic results:

$$n_t \xrightarrow{\mathcal{D}} N \sim \pi_N(n|k) \quad \text{as} \quad t \to \infty \tag{5}$$
$$q_t \xrightarrow{\mathcal{D}} Q \sim \pi_Q(q|k, \hat{q}) \quad \text{as} \quad t \to \infty$$

independent of the starting value q_0. Recall that our goal is to examine the distribution of N, the total number of studies carried out. This asymptotic result tells us that by generating a large enough sample $n_{t+1}, n_{t+2}, \ldots$, we can determine *any* characteristic of $\pi_N(n|k)$ to *any* degree of precision.

Before we can proceed with this algorithm, however, notice that we cannot directly generate q_i values from the conditional distribution of Q in Equation 4. Due to the integral in the denominator, we also cannot find a good, well-behaving approximate distribution that has a calculable (finite) maximum in order to use rejection sampling. We will use the Metropolis method (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953), which generates a value for Q from a ''candidate'' distribution, and accepts that value if it is ''close enough'' that it could have come from the target in Equation 4. We begin with the following candidate distribution:

$$\pi_\Psi(\psi|n, k, \hat{q}) = \psi^{z+\ell}(1-\psi)^{n+b-z-1}/B(z+\ell+1, n+b-z) \quad 0 \le \psi \le 1$$

which matches the power of the $1-\psi$ term with the $1-q$ term in Equation 4. This is a Beta distribution, from which it will be easy to generate samples. The value of ℓ is arbitrary, and can be fine-tuned (for various values of a and b, for example) in order to even more closely approach the target distribution of Q. To correctly simulate Q over its range from ρ to 1, though, we must transform with $\Phi = (1-\rho)\Psi + \rho$, which has the desired range and the following probability density function:

$$\pi_\Phi(\phi|n, k, \hat{q}) = \left(\frac{1}{1-\rho}\right)^{n+b+\ell} (\phi-\rho)^{z+\ell}(1-\phi)^{n+b-z-1}/B(z+\ell+1, n+b-z) \tag{6}$$

As dictated by the Metropolis algorithm, we will generate a q_i from the desired distribution in Equation 4 by applying a decision rule as follows: given a random $u_i \sim \text{Uniform}(0,1)$ and a random ϕ obtained as $\phi = (1-\rho)\psi_{+\rho}$ from a random $\psi \sim \text{Beta}(z+\ell+1, n+b-z)$:

$$\text{if} \quad u_i \le \min\left\{1, \frac{h(\phi)}{h(q_{i-1})}\frac{g^*(q_{i-1})}{g^*(\phi)}\right\} \quad \text{then} \quad q_i = \phi$$
$$\text{if} \quad u_i > \min\left\{1, \frac{h(\phi)}{h(q_{i-1})}\frac{q^*(q_{i-1})}{g^*(\phi)}\right\} \quad \text{then} \quad q_i = q_{i-1}$$

In our model,

$$h(q) = (q-\rho)^z(1-q)^{n+b-z-1}(q)^{k+a-1}$$
$$g^*(q) = \left(\frac{1}{1-\rho}\right)^{n+b+\ell} (q-\rho)^{z+\ell}(1-q)^{n+b-z-1}/B(z+\ell+1, n+b-z)$$

from Equations 4 and 6 respectively. Simplifying, we have:

$$\text{if} \quad u_i \le \min\left\{1, \left(\frac{\phi}{q_{i-1}}\right)^{k+a-1}\left(\frac{q_{i-1}-\rho}{\phi-\rho}\right)^{\ell}\right\} \quad \text{then} \quad q_i = \phi \tag{7}$$

Otherwise, the Metropolis sequence does not move and $q_i = q_{i-1}$. As $i \to \infty$, then the distribution of q_i converges to the desired target distribution. (See Metropolis, et al. (1953) for more details.)

The Gibbs technique is most easily applied here simultaneously with Metropolis sampling as an iterative algorithm. First, generate q_0; subsequently, alternate between generating from the conditional distribution of N (Equation 2) and from the *candidate* conditional distribution of Q (Equation 6). More precisely,

(0.) Generate $q_0 \sim Uniform(0, 1)$

(1.) Generate : n_i from $\pi_N(n|q_{i-1}, k)$

$q_i^* = \phi$ from $\pi_\Phi(\phi|n_i, k, \hat{q})$

(2.) Test q_i^* using the Metropolis criterion specified in Equation 7. Denote an ''acceptable'' value of q_i^* by q_i; otherwise, let $q_i = q_{i-1}$

(3.) Return to (1.) for $i = 1, 2, \ldots, t$.

This combined algorithm also produces two Markov chains, $n_1, n_2, \ldots, n_t$ and $q_1, q_2, \ldots, q_t$, each of which converges in distribution to the desired marginal, as in Equation 5. The issue of how to assess when this convergence happens is still a contentious one among Bayesians. Tanner (1993), Robert (1995), and Cowles and Carlin (1996), among others, summarize several methods to help determine when the Markov chains have reached their equilibrium distributions.

3. Results

3.1. Simulation description

Before applying this theory to a data set, we should assess the behavior of the proposed algorithm. In particular, we want to check the variability of the generated samples as the values for k, q, and $\hat{q}$ vary. The algorithm described in Section 2.3 is run with the number of published studies k set first at 5 and then at 20, in order to see the effect of a small versus a large meta-analysis situation.

The prior parameters on Q (a and b) are each taken to be 5, since this gives the Beta distribution a symmetric bell-shape with somewhat thick tails. A variety of values for $\hat{Q}$ and ρ are chosen; values of $\hat{Q}$ are taken to be 1/10, 1/2, and 9/10 to cover as wide a range as possible. All valid values for ρ are used when $k = 5$; when $k = 20$, simulations are run for only 11 of the possible 33 values for ρ, approximately evenly spaced; see Table 1. Recall that $\hat{Q}$ is constrained by $\hat{Q} = (1 - \rho)z/k + \rho$, where k is given and z must be an integer. The parameter ℓ is (somewhat arbitrarily) set at $k + a - 1$. The power of the q term then equals the power of the $q - \rho$ term in the Metropolis criterion (see Equation 7). As mentioned earlier, this value could easily be adjusted up or down in order to get a higher Metropolis acceptance rate.

We use the GAUSS System (Version 3.01) to produce 10,000 generated numbers of each of N and Q in total, using 10 independent cycles of 1,000 generations each. At the end of the i^{th} cycle of 1,000 generations, $i = 1, 2, \ldots, 10$, several summary values based

Table 1. Parameter values by simulation

$k = 5$				$k = 20$			
Simulation	$\hat{Q}$	ρ	$r = z/k$	Simulation	$\hat{Q}$	ρ	$r = z/k$
1	0.9	0.900	0.0	10	0.9	0.900	0.00
2		0.875	0.2	11		0.867	0.25
3		0.833	0.4	12		0.800	0.50
4		0.750	0.6	13		0.600	0.75
5		0.500	0.8	14		0.000	0.90
6	0.5	0.500	0.0	15	0.5	0.500	0.00
7		0.375	0.2	16		0.333	0.25
8		0.167	0.4	17		0.000	0.50
9	0.1	0.100	0.0	18	0.1	0.100	0.00
				19		0.053	0.05
				20		0.000	0.10

on the conditional Negative Binomial distribution of N are recorded:

- A sample expected value of N: $\hat{E}_i^q[N] = \frac{1}{1{,}000}\sum_{j=1}^{1{,}000}\frac{k}{q_{ij}}$
- A sample variance of N: $\widehat{Var}_i^q[N] = \frac{1}{999}\sum_{j=1}^{1{,}000}\left(\frac{k}{q_{ij}} - \hat{E}_i^q[N]\right)^2$
- An empirical distribution of N: $P_i^q[N = n] = \frac{1}{1{,}000}\sum_{j=1}^{1{,}000}\binom{n-1}{k-1}q_{ij}^k(1-q_{ij})^{n-k}$ for a range of values of n

The superscript q indicates that the value was obtained by averaging across the corresponding conditional values given the q_{ij}'s. These are often called ''Rao-Blackwellized'' estimators, and for variance reasons are better than the usual estimators $\bar{n}_{i\cdot} = \frac{1}{1{,}000}\sum_j n_{ij}$, etc. (see Robert 1990, p. 348). In addition, at the end of the 10 cycles, the following are calculated:

- An overall sample expected value of N: $\hat{E}^q[N] = \frac{1}{10{,}000}\sum_{j=1}^{10}\sum_{i=1}^{1{,}000}\frac{k}{q_{ij}}$
- An overall sample variance: $\widehat{Var}^q[N] = \frac{1}{9}\sum_{i=1}^{10}\left(\hat{E}_i^q[N] - \hat{E}^q[N]\right)^2$

The empirical distributions (frequency histograms) of N are first visually compared across the ten independent cycles to note their stability within each set of parameter values. The graphs are then visually compared across parameter values to note any variability and/or trends.

3.2. Simulation results

Two aspects of the results are under consideration here: (i) behavior of the sample expected values and standard errors; (ii) behavior of the empirical distributions of N. Tables 2 and 3 show the sample expected values and standard errors for the combinations of $\hat{Q}$ and ρ when $k = 5$ and when $k = 20$. Within a value for $\hat{Q}$, we can see that the spread of the distribution, as measured by the standard errors, generally increases as ρ decreases. The trend is more consistent for $\hat{Q} = 1/2$ than for $\hat{Q} = 9/10$ in both tables. Intuitively, this

Table 2. Expected values and standard errors when $k = 5$

Simulation	$\hat{Q}$	ρ	$\hat{E}^q[N]$	$(\widehat{SE}^q[N])$
1	0.9	0.900	5.43	(0.04)
2		0.875	5.49	(0.04)
3		0.833	5.60	(0.04)
4		0.750	5.85	(0.08)
5		0.500	6.64	(0.13)
6	0.5	0.500	8.17	(0.26)
7		0.375	9.04	(0.39)
8		0.167	10.73	(1.47)
9	0.1	0.100	16.72	(3.00)

trend is expected, since a smaller value for ρ indicates that fewer nonsignificant studies are being published. This leads to greater uncertainty in how many unseen studies may have been done, which leads to a distribution on the total number of studies with a larger variance. The distribution on N is not bounded above, but is bounded below, so larger and larger values of N will have larger probabilities of occurring. The expected values will consequently show an increasing trend as well, as is true within every value of $\hat{Q}$ but one (Simulations #18–20). (One erratic iteration of the ten in Simulation #19 enabled much larger values of N to occur.) The trend of increasing standard errors is more consistent within Table 2 than within Table 3. The increase in k, or the values used for ρ, may have led to greater instability in the Metropolis algorithm, perhaps resulting in slower convergence.

When $k = 5$, the empirical distributions of N are *very* stable across the ten cycles within each of the sets of parameter values. (These graphs are not all included here but can be found in Eberly 1994.) Across the values for ρ, but within a value of $\hat{Q}$, the graphs gradually become wider and flatter as ρ decreases, as expected (see Figure 1 for an example). The change is gradual and gives the impression that the estimation is not extremely sensitive to the choice of ρ here. We can see here graphically the numerical trends evident in Table 2: the expected value of N increases slightly as ρ decreases, and the variance of the distribution increases slightly as ρ decreases. When $k = 20$, however, the stability decreases somewhat. Within a set of parameter values, the variability across the ten iterations is

Table 3. Expected values and standard errors when $k = 20$

Simulation	$\hat{Q}$	ρ	$\hat{E}^q[N]$	$(\widehat{SE}^q[N])$
10	0.9	0.900	21.50	(0.06)
11		0.867	21.73	(0.06)
12		0.800	22.18	(0.20)
13		0.600	23.24	(0.55)
14		0.000	26.45	(0.19)
15	0.5	0.500	32.08	(0.65)
16		0.333	37.41	(1.60)
17		0.000	41.35	(0.70)
18	0.1	0.100	85.42	(4.82)
19		0.053	103.00	(19.67)
20		0.000	96.88	(2.78)

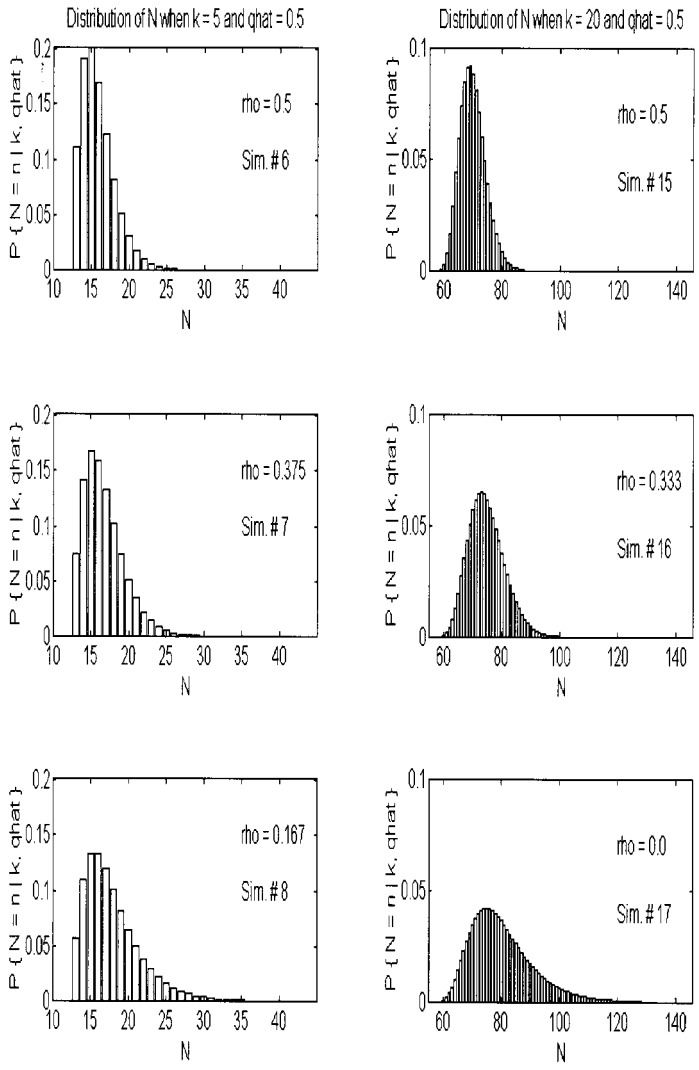

Fig. 1. Simulated distributions of N

occasionally large. (These graphs are also not included.) Across these eleven sets, we see the same trends as when $k = 5$, but the changes in width and height are more dramatic, especially in Simulations #10–17 (see Figure 1 for an example). Within the simulations for $\hat{Q} = 1/10$ (#18–20), however, the changes in the graphs across the values for ρ are very minor. Most likely this is a result of the very narrow range of values for ρ that are possible (0.0–0.1) for these k and $\hat{Q}$ values. Outside of the context of a particular meta-analysis, it is difficult to make more specific conclusions. See Eberly (1994) for more details, and the next section for an example.

4. Application: Lead Exposure and IQ in Children

Needleman and Gatsonis (1990) detail two meta-analyses of studies relating childhood

lead exposure to IQ level. The studies were chosen from the population of all studies on lead exposure and children's neurobehavioral development published since 1972, as found in MEDLINE, meeting programs, and dissertations. Each published study is required by the authors to contain the following in order to be included in a meta-analysis: (i) use of a multiple regression analysis; (ii) a continuous IQ level as the response variable; (iii) lead as a main effect in the regression; (iv) control for non lead covariates in the regression. Twelve studies satisfied these criteria, of which seven measured blood lead and five measured tooth lead. Studies for which all needed information is available give the data found in Table 4 (taken directly from Needleman and Gatsonis 1990, Table 5). It must be noted that neither IQ levels nor lead levels were necessarily measured in the same way across all studies or even within the blood or tooth lead groups.

We assume a one-sided null hypothesis of a positive effect of lead on IQ, i.e., $H_0 : \beta_{lead} \geq 0$, where β_{lead} denotes the regression coefficient. First, we carry out a simple meta-analysis (based on Rosenthal 1978) to obtain an overall Z-value and p-value for the hypothesis of interest:

$$Z_{overall}^{Blood} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} = \frac{-3.86 - 1.67 + \ldots - 1.8}{\sqrt{7}} = -5.35$$

which gives a one-sided p-value of essentially zero. Likewise,

$$Z_{overall}^{Tooth} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} = \frac{-3 - 2.23 + \ldots - 1.17}{\sqrt{5}} = -3.42$$

giving a one-sided p-value of 0.0003. We have to take the original t-values as approximate Z-values here; the sample sizes are large enough that this seems reasonable. Alternatively, we could use the approaches of Fisher or of Mosteller and Bush, as described by Needleman

Table 4. Lead coefficients for full-scale IQ scores

Study	Regression coefficient	Standard error	t-value	Sample size	One-sided p-value
Blood Lead Studies					
Hatzakis, et al.	−0.27	0.07[a]	−3.86[a]	509	0.0001
Hawk et al.	−0.25	0.15	−1.67	75	0.05
Schroeder, et al.	−0.2	0.07[a]	−2.78	104	0.003
Fulton, et al.[b]	−3.7	1.37	−2.77	501	0.003
Yule, et al.[b]	−8.08	4.63	−1.75	129	0.04
Lansdown, et al.[b]	2.15	4.48[a]	0.48	86	0.68
Emhart, et al.	NA[c]	NA	−1.8[a]	80	0.04
Tooth Lead Studies					
Needleman, et al.	−0.21	0.07	−3	218	0.001
Hansen, et al.	−4.27	1.91	−2.23[d]	156	0.01
Winneke, et al.	−0.13	4.66	−0.03[d]	115	0.49
Pocock, et al.[b]	−0.77	0.63	−1.22	388	0.11
Fergusson, et al.[b]	−1.46	1.25	−1.17	724	0.12

[a]Estimated from data in article.
[b]Used log transformation.
[c]Not available.
[d]Obtained from the author.

Table 5. Expected values, standard errors, and credible intervals for N

Simulation	ρ	$\hat{q}$	$\hat{E}^q[N]$	$(\widehat{SE}^q[N])$	C.I.
Blood Lead Studies: $k = 7$ *and* $\hat{r} = 6/7$					
		$a = 5,\quad b = 5$			
1	0.1	0.87	10.86	(2.05	(7,19)
2	0.5	0.93	8.99	(0.79)	(7,13)
3	0.9	0.99	7.31	(0.11)	(7,9)
		$a = 4,\quad b = 2$			
1	0.1	0.87	9.17	(1.57)	(7,16)
2	0.5	0.93	8.14	(0.65)	(7,12)
3	0.9	0.99	7.15	(0.09)	(7,8)
		$a = 1,\quad b = 1$			
1	0.1	0.87	9.17	(1.84)	(7,16)
2	0.5	0.93	8.01	(0.70)	(7,12)
3	0.9	0.99	7.09	(0.08)	(7,8)
Tooth Lead Studies: $k = 5$ *and* $\hat{r} = 2/5$					
		$a = 5,\quad b = 5$			
4	0.1	0.46	11.03	(3.04)	(5,24)
5	0.5	0.70	7.95	(0.85)	(5,14)
6	0.9	0.94	5.42	(0.07)	(5,7)
		$a = 4,\quad b = 2$			
4	0.1	0.46	9.47	(2.64)	(5,20)
5	0.5	0.70	7.44	(0.95)	(5,13)
6	0.9	0.94	5.37	(0.09)	(5,7)
		$a = 1,\quad b = 1$			
4	0.1	0.46	12.72	(5.51)	(5,32)
5	0.5	0.70	7.63	(0.97)	(5,13)
6	0.9	0.94	5.36	(0.10)	(5,7)

and Gatsonis (1990); those approaches yield similar results and will not be reproduced here.

Our p-values are strong indications that the null hypothesis is false, assuming our sample is representative. We run simulations as described in Section 3 in order to make an assessment of the reliability of our results. From the last column in Table 4, six of the seven observed blood lead studies and two of the five tooth lead studies give significant results at $\alpha = 0.05$. Hence, the Gibbs sampler will be run first with $k = 7$ and $\hat{r} = 6/7$, and second with $k = 5$ and $\hat{r} = 2/5$. The program is run to produce 5,000 generated numbers of each of N and Q in total. By Equation 1, then, we can choose several values for ρ and calculate the corresponding values for $\hat{q}$. In order to capture any trend as the value of ρ changes, we will take $\rho = 1/10$, $1/2$, and $9/10$. The simulations are run for three sets of prior parameters for Q: $a = b = 5$ (a bell-shaped density for Q), $a = 4$ and $b = 2$ (a skewed density with mean equal to 2/3), and $a = b = 1$ (a uniform density). The value for ℓ is adjusted (up or down, as necessary) from its initial setting at $k + a - 1$ to ensure that the Metropolis sampling accepts at least 75% of the generated candidate values.

At the end of the 5,000 cycles, the following are calculated: $\hat{E}^q[N]$, $\widehat{Var}^q[N]$, and a 95% credible interval for N. A Bayesian $1 - \gamma$ credible interval (N_ℓ, N_u) is calculated from $\int_{N_\ell}^{N_u} \pi_N(n|k)dn = 1 - \gamma$. The interval has the intuitive interpretation

that $P(N_\ell \le N \le N_u|k) = 1 - \gamma$. Here we choose to use an equal-tailed interval with $\gamma = 0.05$, so then $\int_{-\infty}^{N_\ell} \pi_N(n|k)dn = \int_{N_u}^{\infty} \pi_N(n|k)dn = \gamma/2$. This can be calculated by using the sample quantiles: rank the n_i-values and denote them as $n_{(1)} \le n_{(2)} \le \ldots \le n_{(5{,}000)}$. Take $n_{(5{,}000.\gamma/2]+1)}$ as the estimate of N_ℓ, and $n_{((5{,}000.(1-\gamma/2))+1)}$ as the estimate of N_u. (Here, (y) denotes the largest integer less than or equal to y. If $(5{,}000(\gamma/2))$ or $(5{,}000(1-\gamma/2))$ is itself an integer, then 1 is not added.) This interval is equivalent to that obtained by inverting the empirical CDF at $\gamma/2$ and at $1-\gamma/2$. The results are shown in Table 5.

It is clear (and reassuring) that the results are very consistent across the various values chosen for a and b. Given the assumptions made about the prior distribution on Q, these results tell us that there could be about $(7-7=)0$ to $(11-7=)4$ blood lead studies on this hypothesis which were unseen. The researcher must now make his or her best guess at an appropriate value for ρ. In the most optimistic case, $\rho = 9/10$; *most* nonsignificant and *all* significant studies are published. In this case, we expect no unseen studies, so our sample of published studies can be considered entirely trustworthy. In the least optimistic case, $\rho = 1/10$ and *most* nonsignificant studies are *not* published, whereas all significant studies are. In this case, we could have four unseen studies. If all of them are *strongly* nonsignificant, or significant in the opposite direction, it is possible that our combined p-value could be overturned. However, Needleman and Gatsonis (1990, p.677) make a very good point: ''Given the expense of conducting human studies of lead exposure and the amount of attention directed to this question, it is unlikely that this number of negative studies have escaped notice.'' For the tooth lead studies, there could be about $(5-5=)0$ to $(13-5=)8$ unseen studies. As above, in the most optimistic case, we expect no unseen studies, and the results of our meta-analysis seem trustworthy. In the least optimistic case, there could be more nonsignificant studies out there than studies on hand. The meta-analysis could be giving us very biased results. Again, though, it seems unlikely that results with strong conclusions contrary to published conclusions would not have been noticed. Clearly, knowledge of the subject matter is needed to make a judgment on the probable value for ρ.

In cases where individual study p- and Z-values are available, it may be helpful to compare the simulation results to two other source augmentation methods. Rosenthal's fail-safe (*FS*) number (Rosenthal 1979) calculates the number of unseen studies *averaging null results* (i.e., a p-value of 0.5 or a Z-value of zero) needed to bring a significant overall p-value to a specified level. The fail-safe numbers are based on the same method of combining Z-values that was used above. Since those two Z-values are both significant, it makes sense to calculate Rosenthal's estimates and compare them with our simulation results:

$$FS_{Blood} = \left(\frac{\left(\sum_{i=1}^{k} Z_i\right)^2}{(1.645)^2} - k \right)^+ = \left(\frac{(-3.86 - 1.67 + \ldots - 1.8)^2}{(1.645)^2} - 7 \right)^+ = 66.99$$

$$FS_{Tooth} = \left(\frac{\left(\sum_{i=1}^{k} Z_i\right)^2}{(1.645)^2} - k \right)^+ = \left(\frac{(-3 - 2.23 + \ldots - 1.17)^2}{(1.645)^2} - 5 \right)^+ = 16.63$$

Hence, 66 unseen studies giving null results are needed to overturn this combined p-value of zero from the blood lead studies, while 16 are needed to overturn the 0.0003 from the tooth lead studies. Since from Table 5 only 5 to 11 studies of any kind

(significant or not, published or not, measuring blood or tooth lead levels) are expected to be out there on average, it seems highly improbable that there are enough unseen null studies to overturn the p-value, no matter what the value of ρ.

An additional comparison can be made with Gleser and Olkin's (1996) frequentist method of estimating N. This method assumes that we have obtained the m smallest p-values, plus a random sample of $k - m$ p-values from the remaining $N - m$ studies out there. The value of m can be estimated empirically be plotting the ranked p-values $p_{(i)}$ versus i for the k observed studies. These data should be roughly well-fitted by two straight lines, one through the first m points and another through the remaining $k - m$ points. Plotting the ranked p-values $p_{(i)}$ versus i from Table 4 indicates that $m = 6$ for the blood lead and $m = 4$ for the tooth lead studies (plots are not shown here). Thus, $\hat{N}_{Blood} = (m - 1)/p_{(m)} = 5/0.05 = 100$ and $\hat{N}_{Lead} = 3/0.12 = 25$. Unfortunately, having to estimate m in this manner means $\hat{N}$ is no longer unbiased. Gleser and Olkin (1996) also give a formula for a $100(1 - \alpha)\%$ lower bound for N as $\min_{q \geq 0}\{q : F_{2m,2(q+1);\alpha} < (q + 1)(1 - p_{(m)})/(mp_{(m)})\} + m$. We take $\alpha = 0.025$ to match the level of our credible intervals, giving $N_L^{Blood} = 8$ and $N_L^{Lead} = 7$. These lower bounds are very close to our lower bounds.

Taken in concert, the three source augmentation methods discussed here offer reassurance that the meta-analyses are reliable. Rosenthal (1979) offers his own guidelines on what an ''unlikely'' number of unseen studies might be. He suggests that some fields may consider 100 or 500 unseen studies plausible, whereas other fields may deem only 10 or 20 as likely. Rosenthal's recommendation is to consider $5k + 10$ the level at which the number of unseen studies becomes implausible. The $5k$ suggests that it is unlikely that there are more than five times as many studies filed away as there are on hand, while 10 sets the minimum number of studies at 15 when $k = 1$. In this example, the cutoffs would be 45 and 35 for the blood and tooth studies, respectively.

As a caution, the p-values calculated in Equation 4 above are based on what may or may not be a good estimate of the overall Z-values. One must always keep in mind that there are many other ways to calculate an overall p-value, ones that, for example, take sample sizes or sample variances into account (see Rosenthal 1978). Some of those methods could give nonsignificant overall results, in which case any consideration of *FS* is nonintuitive. In addition, since this is a one-sided hypothesis testing situation, the researcher must consider the possibility of unpublished studies that are significant in the *opposite* direction. Rosenthal's estimates are a useful (and possibly reassuring) comparison to make when the data are available to calculate them. However, they are strictly ad hoc estimates and the statistical properties associated with them are not known; caution should be used in interpreting them.

5. Conclusions

We have derived a method for approximating the total number of studies done on a particular hypothesis, given a selection probability (ρ), a distribution of the probability of publication (Q), and a meta-analysis of k available studies. The theory is complex only in that it must adapt to circumvect practical computational difficulties (i.e., Metropolis simulation and Gibbs sampling). One drawback of this theory, of course, is that the prior distribution on Q must be specified. Very few researchers will be able to choose parameter

values for the Beta prior distribution with any degree of assuredness. Our application shows, however, that those choices do not much influence the results of the simulations. Research in Bayesian statistics has shown that (as we saw here) posterior distributions can be robust to the choice of the prior distribution; see, for example, Berger (1993). In addition, there is a small but growing literature on the formal elicitation of prior information from expert opinion, which may be well-suited for this situation (see for example, Carlin, Chaloner, Church, Louis, Matts 1993, or Steffey 1992). Further investigation should be done regarding the effect of varying a and b on the stability of the simulations and on the precision of the approximations.

Another potential problem is the violation of assumptions. It is conceivable that the probability of publication is *not* constant across studies. In situations where a great deal of funding is allocated for large-scale nationwide clinical trials, for example, it is almost a certainty that these results will be published, significant or not. As a related issue, the biggest criticism of the fail-safe numbers is that they fail to distinguish between studies which are significant due to a large effect size, and studies which are significant due to a large sample size. Our method sidesteps this criticism by requiring that all studies considered are roughly of the same size. In practice, this requirement is not likely to be satisfied, but it is likely to be an improvement over the fail-safe criterion. As a partial fix, the next step would be to place a prior distribution on the probability r, thus allowing it to vary across studies.

Given the simulation program, these methods are easy to implement and easy to interpret. An obstacle to using these methods in a specific application is that a value (or possibly values) for ρ must be chosen. A researcher must have a good familiarity with both the publication process and the activities of other researchers in his or her field to be able to give a reliable estimate. We recommend, therefore, that the simulations always be run for a range of values for ρ. Hopefully, from personal experience, this range can at least be limited to only a small portion of the interval (0,1). The application of this theory would be much improved if a method for estimating ρ were developed. Another disadvantage is that any application of this theory can only start from a count of the number of significant studies (i.e., to calculate z/k), not from individual p-values nor Z-values. It seems there is a loss of information at some level here. Sample sizes and sample variances from the studies under consideration do not affect this procedure, when ideally it seems they should. The next step is perhaps to consider a model that depends not only on ρ, but also on other relevant covariates. Either ρ could be modeled deterministically, by choosing some function of the covariates, or a prior distribution for ρ could be chosen. A further and perhaps more realistic generalization of another aspect of the model would be to let $Q = \delta R + \rho(1 - R)$, so that not all significant studies are assumed published. In conclusion, using a range of values for ρ and the Gibbs/Metropolis procedure, a reasonable picture of the number of unseen studies can be formed for a specific meta-analysis application. Rosenthal's fail-safe estimates and Gleser and Olkin's lower bounds can be used as comparative indications of the reliability of a significant overall p-value.

6. References

Bayarri, M.J. and DeGroot, M. (1986). Bayesian Analysis of Selection Models. Technical Report 365, Dept. Statistics, Carnegie-Mellon University.

Bayarri, M.J. and DeGroot, M. (1991). The Analysis of Published Significant Results. Technical Report 91-21, Dept. Statistics, Carnegie-Mellon University.

Begg, C.B. and Berlin, J.A. (1988). Publication Bias: A Problem in Interpreting Medical Data. Journal of the Royal Statistical Society – Series A, 151, 419–463.

Begg, C.B. and Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias. Biometrics, 50, 1088–1101.

Berger, J. (1993). An Overview of Robust Bayesian Analysis. Test, 3, 5–124.

Berlin, J.A., Begg, C.B., and Louis, T.A. (1989). An Assessment of Publication Bias Using a Sample of Published Clinical Trials. Journal of the American Statistical Association, 84, 381–392.

Carlin, B.P., Chaloner, K., Church, T., Louis, T.A., and Matts, J.P. (1993). Bayesian Approaches for Monitoring Clinical Trials With an Application to Toxoplasmic Encephalitis Prophylaxis. The Statistician, 42, 355–367.

Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler. The American Statistician, 46, 167–174.

Cleary, R.J. and Casella, G. (1997). An Application of Gibbs Sampling to Estimation in Meta-Analysis: Accounting for Publication Bias. Journal of Educational and Behavioral Statistics, 22, 141–154.

Cowles, M.K. and Carlin, B.P. (1996). MCMC Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association, 91, 883–904.

Dear, K.B.G. and Begg, C.B. (1992). An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis. Statistical Science, 7, 237–245.

Dickersin, K., Min, Y.-I., and Meinert, C.L. (1992). Factors Influencing Publication of Research Results. Journal of the American Medical Association, 267(3), 374–378.

Easterbrook, P.J., Berlin, J.A., Gopalan, R., and Matthews, D.R. (1991). Publication Bias in Clinical Research. Lancet, 337, 867–872.

Eberly, L.E. (1994). Estimating the Number of Unseen Studies in a Meta-Analysis. M.S. Thesis, Biometrics Unit, Cornell University, Ithaca, NY.

Frongillo, E. (1991). Combining Information Using Hierarchical Models. Ph.D. Dissertation, Biometrics Unit, Cornell University, Ithaca, NY.

Gelfand, A.E. and Smith, A.F. (1990. Sampling-based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association, 85, 398–409.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721–741.

Glass, G.V., McGaw, B., and Smith, M.L. (1981). Meta-Analysis in Social Research. Beverly Hills, CA: Sage Publications.

Gleser, L.J. and Olkin, I. (1996). Models for Estimating the Number of Unpublished Studies. Statistics in Medicine, 15, 2493–2507.

Hedges, L.V. (1984). Estimation of Effect Size under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences. Journal of Educational Statistics, 9, 61–85.

Hedges, L.V. (1992). Modeling Publication Selection Effects in Meta-analysis. Statistical Science, 7, 246–255.

Iyengar, S. and Greenhouse, J.B. (1988). Selection Models and the File Drawer Problem. Statistical Science, 3, 109–135.

Light, R. and Pillemer, D. (1984). Summing Up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21, 1087–1091.

National Research Council (1992). Combining Information: Statistical Issues and Opportunities for Research. National Academy Press: Washington, D.C.

Needleman, H.L. and Gatsonis, C.A. (1990). Low-level Lead Exposure and the IQ of Children. Journal of the American Medical Association, 263, 673–678.

Robert, C.P. (1990). The Bayesian Choice: A Decision-Theoretic Motivation. New York: Springer-Verlag.

Robert, C.P. (1995). Convergence Control Methods for Markov Chain Monte Carlo Algorithms. Statistical Science, 10, 231–253.

Rosenthal, R. (1978). Combining Results of Independent Studies. Psychological Bulletin, 85, 185–193.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. Psychological Bulletin, 86, 638–641.

Silliman, N.P. (1997). Hierarchical Selection Models with Applications in Meta-Analysis. Journal of the American Medical Association, 92, 926–936.

Steffey, D. (1992). Hierarchical Bayesian Modeling with Elicited Prior Information. Communications in Statistics A, 21, 799–821.

Tanner, M.A. (1993). Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 2nd ed. New York: Springer-Verlag.

Tanner, M.A. and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82, 528–540.

White, K.R. (1982). The Relation Between Socioeconomic Status and Academic Achievement. Psychological Bulletin, 91, 461–481.

Received May 1997
Revised June 1998