Slamming the sham: A Bayesian model for adaptive adjustment with noisy control data*

Andrew Gelman† and Matthijs Vákár‡

27 Jan 2020

Abstract

It is not always clear how to adjust for control data in causal inference, balancing the goals of reducing bias and variance. In a setting with repeated experiments, Bayesian hierarchical modeling yields an adaptive procedure that uses the data to determine how much adjustment to perform. We demonstrate this procedure on two real examples, as well as on a series of simulated datasets. We also discuss the relevance of this work to causal inference and statistical design and analysis more generally.

1. Introduction

Consider the following problem. Experiments $j = 1, \ldots, J$ are performed, and each is paired with a sham experiment with a null treatment. Label the estimated treatment effects for each experiment j as y_{j1} for the active data and y_{j0} for the sham data. It is standard practice to estimate the treatment effect in experiment j as $y_{j1} - y_{j0}$. But this bias adjustment can add noise. In many cases, it is not a priori obvious whether the sham experiments can be safely discarded or not. The same problem arises in observational studies in economics with the difference-in-differences estimate (see, e.g., Ashenfelter, Zimmerman, and Levine, 2003), where, again, subtracting the baseline difference can reduce bias but at the cost of increasing variance.

How can we decide whether to adjust for the sham data and how best to do so, if we do adjust? We propose a hierarchical Bayesian approach which is broadly consistent with modern ideas of regularization in causal inference for varying treatment effects (e.g., Hill, 2011, and Wager and Athey, 2018). We move beyond standard Bayesian meta-analysis (e.g., Smith, Spiegelhalter, and Thomas, 1995, and Higgins and Whitehead, 1996) by partially pooling biases as well as the treatment effects, thus allowing an adjustment that adapts to observed variation in the sham data. Our approach uses Bayesian multilevel modeling and is most generally effective when the number of experiments, J, is large. When J is small, strong prior information is required to most efficiently use the sham information; when J is large, the relevant hyperparameters can be estimated from the data.

The core contributions of this paper are: (1) a focus on a problem that arises in many areas of science when analyzing repeated controlled experiments, and where an existing default method can yield demonstrably poor performance, (2) a solution, along with code to implement it, (3) a method for interpreting the resulting estimate as an approximate partial adjustment, and (4) a set of simulation-based evaluations of the method that are directly relevant to the ways in which these studies are reported.

We demonstrate the need for a solution to the sham-adjustment problem, and our recommended method, in the context of two real examples. First, we consider a series of laboratory experiments on the effects of electromagnetic fields on calcium flow in the brain. The results of these experiments were influential in a public health debate regarding cancer clusters that had been found near electric power lines. For the experimental results under study, it possible to greatly improve the published analysis by modeling the bias rather than simply subtracting the sham estimate, and the new

*We thank the U.S. Office of Naval Research, Institute for Education Sciences, and Sloan Foundation for partial support of this work. Data and code are at https://github.com/VMatthijs/Slamming-the-sham

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡Department of Information and Computing Sciences, Utrecht University, Utrecht.

analysis alters the scientific conclusions. As a second case study, we consider a recent, highly cited meta-analysis on repetitive transcranial magnetic stimulation as a treatment for depression by Berlim et al. (2014). We emphasize that our method addresses a problem that is widely found across science, however, whenever repeated controlled experiments are analyzed, and we could have equally have considered the studies of Fuchikami et al. (2010), Kádár et al. (2011), or Le Quément et al. (2012), to name a few.

2. The problem and proposed solution

2.1. The model and two estimates

Suppose that, for each of $j = 1, \ldots, J$, two experiments have been conducted, yielding estimate y_{j1} and standard error s_{j1} from the active-treatment experiment and estimate y_{j0} and standard error s_{j0} from the sham-treatment experiment. We assume the following model that is intended to capture the experiment and estimation process for each pair of experiments j, where we assume statistical independence between all pairs of experiments:

$$\begin{aligned} y_{j1} &\sim \text{normal}(\theta_j + b_j, s_{j1}) \\ y_{j0} &\sim \text{normal}(b_j, s_{j0}). \end{aligned} \tag{1}$$

Here, θ_j is the treatment effect of interest and b_j is an experimental bias shared by the real and sham treatments. In modeling the bias in this way we are following the general approach of Greenland (2005). For simplicity of presentation and for application to the meta-analysis problem, we shall assume that the estimates and standard errors are given, and that the sample size in each experiment is large enough that it is reasonable to approximate the information from the data in the form of normal likelihoods with known variances. It would not materially affect the methods or conclusions of this paper if we were to go to the raw data (where available) or to replace the normal with t likelihoods corresponding to the degrees of freedom of the data in each experiment.

We start by considering two estimates of the treatment effect θ_j: the exposed-only estimate, y_{j1}, or the difference estimate, $y_{j1} - y_{j0}$.

Under model (1), the difference estimate is unbiased—indeed, it is the only unbiased estimate of θ_j. However, performing this subtraction adds noise, doubling the variance if the standard errors of the active treatment data and sham are the same. If the bias b_j in the experiments were zero, the exposed-only estimate would clearly be the better choice. More generally, depending on the size of the bias, b_j, it could be more effective to partially adjust for the sham rather than to fully subtract y_{j0}.

At this point, a scientist might feel that the safe choice would be to use the difference estimate, paying the price of a higher mean squared error, as it could seem risky to accept bias. Researchers are often trained to think of bias as the primary concern, with the minimum-variance unbiased estimator being optimal (Lehmann and Scheffe, 1950). We suspect that such an attitude is not as prevalent as in the past, now that we are used to regularization in methods ranging from lasso to deep learning to multilevel regression, but it remains a starting point in many analyses.

In the present paper we shall consider the exposed-only and difference estimates as two extreme cases of a Bayesian procedure that performs meta-analysis on the treatment effects and biases. We first present the Bayesian model, then demonstrate its merits on the applied example that motivated this research as well as on a more recent example that illustrates different aspects of the model, and then present methods for understanding and evaluating the inferences.

2.2. Multilevel model and Bayesian analysis

We have set up model (1) in a way to reflect the scientific choices indicated in design and data collection. The next step is the model for the treatment effects and the biases. This is the multilevel part of the model, and by default we will use normal distributions (again, assuming independence between different values of j):

$$\begin{aligned} b_j &\sim \text{normal}(\mu^b, \sigma^b) \\ \theta_j &\sim \text{normal}(\mu^\theta, \sigma^\theta). \end{aligned} \tag{2}$$

We briefly go through the hyperparameters of this model:

- $\mu^\theta, \sigma^\theta$ are the mean and standard deviation of the true effects. μ^θ and σ^θ determine the partial pooling in the estimates of the individual θ_j's.
- μ^b is the average experimental bias and will equal zero if the sham treatments have no effect.
- σ^b is the variation in the biases across experiments and, again, will equal zero if the sham treatments have no effect.

We need to include an average sham effect and variation in the sham effects in the model to allow for the possibility of bias. This is a matter of respecting the experimental design: the sham treatments were included in the study for a reason.

We can fit the model using Bayesian inference with default uniform priors on the hyperparameters $\mu^\theta, \sigma^\theta, \mu^b, \sigma^b$, with the understanding that informative priors could be used in problems where such prior information is readily available. We choose a Bayesian approach (rather than using marginal maximum likelihood to obtain a point estimate of the hyperparameters) because it accounts for the uncertainty in the hyperparameters, and also for computational convenience—we can fit our model directly in Stan (Stan Development Team, 2012), and it is easy to extend the Stan model to include departures from normality, linearity, and exchangeability as desired.

Model (2) represents a default, or starting point. In real-world meta-analyses the J studies will differ in various known ways. Suppose we have a predictor x_j assigned or observed for each study, j. Then it will make sense to allow the expected treatment effect to vary by x, thus replacing the exchangeable model for θ in (2) by something like,

$$\theta_j = g(x_j), \tag{3}$$

where g is a stochastic function whose distribution will itself depend on hyperparameters, for example a linear regression with errors, $g(x_j) \sim \text{normal}(a + bx_j, \sigma^\theta)$, or a Gaussian process that penalizes discrepancies between $g(x_j)$ and $g(x_k)$ for nearby pairs (x_j, x_k). The choice of model for g will depend on the particular applied problem.

It would also be possible to add structure to the model for the biases b_j. For example, a correlation between b_j and θ_j would allow biases to be larger under conditions of larger treatment effects, which could make sense in some contexts.

2.3. Frequency evaluation

In applied statistics it is not enough to come up with a good estimate; it is also necessary to understand it and compare to previously existing approches. To this end, we compare the estimated treatment effects $\hat{\theta}_j$ under the hierarchical model to the exposed-only estimates, y_{j1}, and the difference estimates, $y_{j1} - y_{j0}$. We conduct this evaluation using a simulation study, as we

demonstrate in Section 5 for our motivating example. The simulation study is conducted to allow a range of values for the crucial parameter σ^b which governs the value of the information from the sham experiments.

For each of the three estimates, we then compute the following four summaries: (i) the proportion of the J estimates that are statistically significant (that is, where the estimate ± 1.96 standard errors or Bayesian 95% posterior interval excludes zero), (ii) the type S error rate (the proportion of statistically significant estimates that are the wrong sign), (iii) the mean squared error of the J estimates compared the true values θ_j (which by the design of the simulation are known to us), and (iv) the correlation between the ranks of the J estimates and the ranks of the true θ_j's.

We choose these summaries because they represent four different practical goals of this sort of study: (i) identification of experiments where the treatment effect is statistically distinguishable from zero, (ii) validity of these claims of confidence, (iii) accurate estimation of treatment effects, and (iv) ranking of which results are strongest and most worthy of further study. It is important in any frequency evaluation to consider statistical properties that are relevant to the task at hand, and we argue for the relevance of these measures in the context of our applied example.

2.4. Relevance and novelty of this procedure

This model can apply to a large set of problems of repeated controlled experiments, such as arise in biology, medicine, policy analysis, and other fields where a treatment effect is conjectured to vary in some unknown way as a function of input conditions, so that the point of the study is not merely to estimate an average treatment effect but also to estimate the individual θ_j's. In Section 3, we consider an example from biology in which the goal was to estimate the dependence of θ on x; in Section 4 we consider a medical example where the distribution of the θ_j's was of interest.

The hierarchical model and Bayesian computation used in this paper are now familiar statistical tools. What is new here is, first, their application to a causal inference setting where it is often standard practice to simply subtract sham estimates (sometimes called a difference-in-difference procedure) rather than to jointly model active and sham data; and, second, the frequency evaluation demonstrating the superiority of the modeling approach under a wide range of conditions; and, third, the expression of the Bayesian estimate as an approximate fractional adjustment for the sham, which links these results to existing practice.

3. Applied example 1: Magnetic fields and calcium efflux

3.1. Background

The 1980s saw a concern regarding health effects of low-frequency magnetic fields, as a result of some findings in epidemiology that children living near electric power lines had elevated risks of leukemia, and this caught the interest of the news media (Brodeur, 1989a, 1989b, 2000). One posited mechanism for a carcinogenic effect here was that magnetic fields interfered with cell structure, and this general model was studied in a series of experiments conducted at the U.S. Environmental Protection Agency, measuring the effects on calcium efflux in chick brains. The studies were carefully conducted with an eye toward theory, measurement, and statistical design (Blackman, 2015). Each chick brain was divided in two, with one half of the brain randomly assigned to the treatment of exposure to an alternating current magnetic field at a specified frequency and the other brain half given the control of no exposure to the field. Between 28 and 36 chicks were employed in each experiment, and 38 experiments were performed, representing magnetic field frequencies ranging from 1 to 510 Hz; see Blackman et al. (1988).

		Sham treatment		Real exposure
Frequency (Hz)	n	Estimate y_{j0} (s.e. s_{j0})	n	Estimate y_{j1} (s.e. s_{j1})
1	32	−0.005 (0.041)	32	0.036 (0.041)
15	32	0.013 (0.042)	36	0.173 (0.034)
30	32	0.033 (0.032)	32	0.107 (0.035)
45	32	−0.010 (0.032)	32	0.181 (0.052)
...	...	...	...	...

Table 1: *A portion of the data summaries from the chick brains experiment reported by Blackman et al. (1988). Data continue at 15 Hz intervals all the way up through 510 Hz. As can be seen from the above numbers, the sham estimates are statistically indistinguishable from zero, whereas the effects are clearly positive for many of the real experiments.*

As a check against systematic bias, each experiment was repeated under "sham" conditions, with the same setup but with the magnetic field turned off. Each sham and real experiment was then analyzed to produce an estimated relative effect, along with a standard error. The experimental design also included clustering, but we do not further consider that here. Unfortunately the authors refused to share their data when requested, and so in our analysis we are restricted to the published data summaries, which are the estimates and standard errors for each sham and real experiment. A subset of these data summaries are displayed for clarity in Table 1.

In the published analysis, the effect of magnetic fields at each frequency was estimated by subtracting the estimates from the real and sham exposures, adding the variances as is appropriate for independent experiments. The estimate for each experiment j is then $y_{j1} - y_{j0}$, with standard error $(\sigma_{j1}^2 + \sigma_{j0}^2)^{1/2}$.

Is it appropriate to subtract the sham estimate? An alternative would be to simply use the estimate from the real exposure, y_{j1} with its standard error, σ_{j1}, which discards the sham data entirely and has the benefit of having approximately half the variance of the differenced estimator.

The difference, $y_{j1} - y_{j0}$, would typically be considered a safe and conservative estimate as it corrects for any biases shared by the two experiments, and it indeed was used in the published paper and not questioned in that literature. However, as we shall see in our discussion of the inferences and conclusions drawn from these data, reliance on the noisy differenced estimator may well incur real scientific costs.

3.2. Originally published analysis

Blackman et al. (1988) presented the differenced estimates and categorized them based on levels of statistical significance relative to the hypothesis of zero effects. The top row of Figure 1a shows redrawn versions of the graphs in that paper. The top-left graph displays point estimates, shading those that are statistically significant. The top-right graph shows p-values of the hypothesis of zero effect at each frequency.[1] The authors divide these into three categories: those with p-values less than 0.01, those with p-values between 0.01 and 0.05, and the rest.

This division based on statistical significance was a mistake, and it is a common mistake in applied statistics; see Gelman and Stern (2006). Seemingly major differences in p-values are not necessarily statistically significant or even close to significant. For example, p-values of 0.20 and 0.01 correspond to z-scores of 1.28 and 2.33, respectively (using the normal distribution here for

[1]Our Figure 1b is slightly different from Figure 2 of Blackman et al. (1988) for reasons that are not clear to us, as our displayed p-values are consistent with those in Table 1 of Blackman et al., but in any case the differences are minor and do not affect the arguments of this paper.

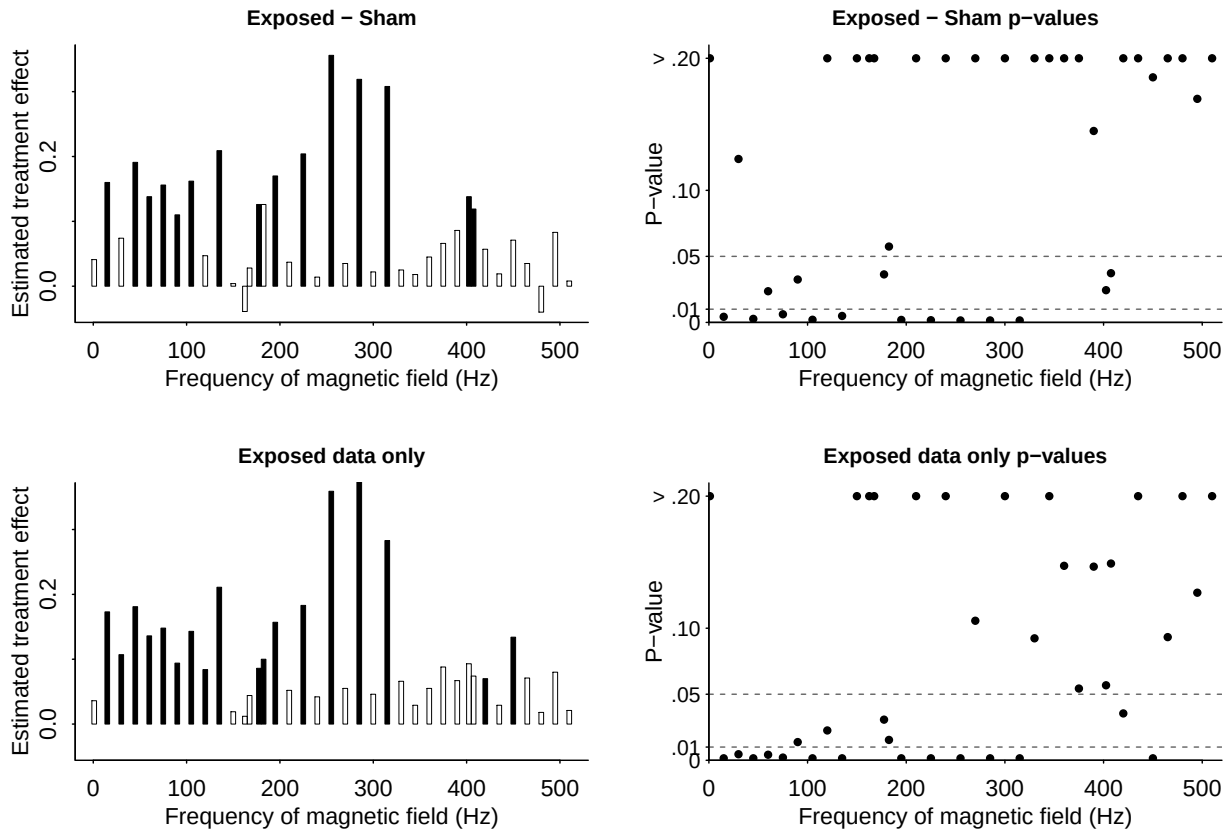

Figure 1: Top row: *Redrawn versions of the graphs of Blackman et al. (1988), summarizing the chick brains data by categorizing estimates at different frequencies based on their statistical significance: (a) Estimates $y_{j1} - y_{j0}$ plotted vs. frequencies x_j. For three of the frequencies (165, 180, and 405 Hz) the experiment was performed twice, and in these cases we have jittered the two experimental results so they both appear on the graph. Each bar is shaded if the experimental result is statistically significant at the 5% level based on the appropriate t distribution. (b) Results of each experiment displayed as a p-value.*
Bottom row: *Corresponding plots using only the exposed data, y_{j1}. The patterns are similar but with enough differences to change some of the reported results.*

simplicity). So, even though $p = 0.20$ seems like no evidence at all, while $p = 0.01$ appears to be a very strong result, their difference is a mere 1.05 standard errors, which can easily occur by chance.

The use of a p-value-based decision rule had consequences. In the paper under discussion, Blackman et al. (1988) used the summary shown in the top-right graph of Figure 1 to draw the following conclusions: "those data with P-values less than 0.01, which extend from 15 to 315 Hz, could form one set composed of two groups of 30 Hz . . . the response at 60, 90 and 180 Hz, the first odd multiple of 60 Hz, with an elevated but not statistically reliable response at 30 Hz, may be part of a second set . . . the response at 405 Hz may represent still another set . . . ." To their credit, the authors emphasized that these are "only hypothetical constructs," but these noisy results form the empirical conclusions of the paper and they motivate in the published paper a further three-page speculation about physical models.

The specific claims from these and similar experiments also influenced researchers' perceptions of underlying mechanisms. For example, Brodeur (1989b) reports, "Blackman was trying to figure

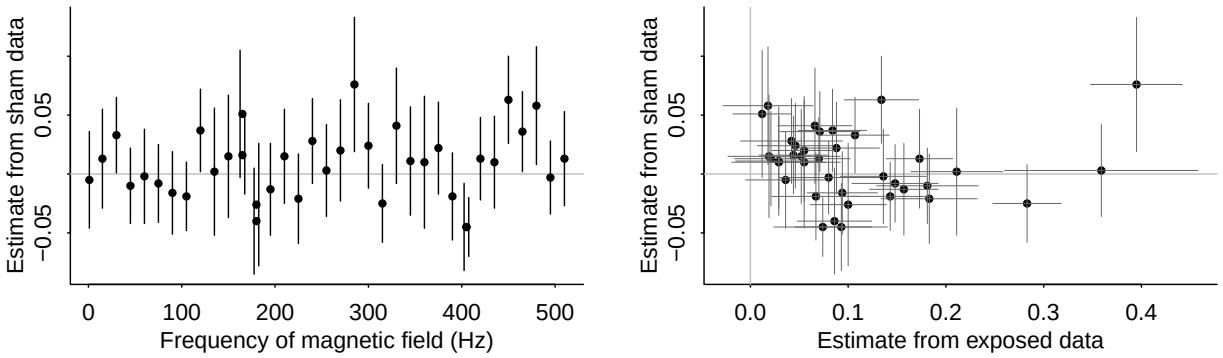

Figure 2: *(a) Estimates ± standard errors of the effect of the sham treatment as a function of frequency of the (turned off) electromagnetic field; (b) Sham vs. exposed estimates. Unsurprisingly, given the careful design of the experiment, there is no evidence that the sham effects are anything other than zero.*

out why fields with frequencies of fifteen, forty-five, seventy-five, and a hundred and five hertz should have such a strong effect on calcium-ion outflow from chick-brain tissue, while fields of thirty, sixty, and ninety hertz produced only a weak effect."

3.3. Exploration of the sham data

For now let us set aside concerns about summarizing experimental results by discretizing p-values, an approach that has been increasingly contested in recent years (Wasserstein and Lazar, 2016), and instead focus on the question of what should be done with the data from the sham experiments in the chick study.

A glance at Table 1 suggests that nothing much seems to be going on in the sham data, which is confirmed by examination of the entire dataset: the estimates fluctuate around the zero, with the amount of variation consistent with the reported standard errors; see Figure 2a. This impression can be confirmed with a simple χ^2 test: $\sum_{j=1}^{38}(y_{j0}/\sigma_{j0})^2 = 21.3$, which is quite a bit *less* than would be expected under the χ^2_{38} distribution. This suggests there may be an problem with the standard errors, as they seem to be too conservative—perhaps there was an error in their computation, as the data were collected using a clustered design and perhaps this was not correctly handled in the standard error calculations—but, in any case, there is no evidence for any variation in the effects of the sham treatment. Furthermore, the mean of the 38 sham estimates is 0.01, which is both substantively and statistically insignificantly different from the null, so the data do not contradict the model of no sham effect. This should be no surprise—given that the experiment was conducted by experts in the field, we would not expect a null treatment to have any effect, and the sham experiments represent an abundance of caution more than anything else.

In our remaining treatment of these data we shall take the sham estimates and standard errors as reported; arguably, though, it would make sense to scale all the standard errors down by a factor of $\sqrt{21.3/38}$ as an approximation to the adjustment that would be required, under the assumption that some mistake was made in their calculation. Scaling these standard errors down would not affect our main conclusions; indeed it would just make our advocacy of an alternative analysis even stronger by increasing the precision of our inferences.

To continue with our main thread, in Figure 2b we look for patterns in the sham data another way, by plotting the sham estimate y_{j0} vs. the exposed estimate y_{j1} for each frequency j. We see

no pattern, which again is consistent with the sham estimates being pure noise.

3.4. Analysis not adjusting for the sham data

If the sham estimates are indeed nothing but noise, then it makes sense not to include them in the estimated treatment effects. The resulting unadjusted analysis is simple: just report y_{j1} with standard error σ_{j1} at each frequency j. We could almost describe this as "analysis ignoring the sham data" but that would not quite be correct. We did not ignore the sham data: we only decided to exclude the sham data from our inferences after first analyzing the sham results and finding no evidence distinguishing them from pure noise.

The bottom row of Figure 1 shows the results. We use the same sorts of displays as used in the earlier published paper, not because we think it appropriate to summarize a set of experiments using statistical significance but because we wish to demonstrate the potential practical gains that could come from switching to the undifferenced estimates, even without considering alternative inferential summaries.

For this example, it is clear from a modern perspective that the estimates $y_{j1} - y_{j0}$ are inferior to the simple y_{j1}. The sham experiments may well have been an important part of the design of the study, as they rule out a potential threat to validity in the causal inferences, but given what the data look like, it is not necessary to include their data in the final estimates.

The challenges we address in this paper are, first, to come to this conclusion in a more systematic way; second, to situate this in a general framework that can apply to other designs; third, to come up with a compromise solution for settings where the sham data are noisy but contain some information; and, fourth, to be able to report such a compromise estimate in a reasonable way.

3.5. Scientific consequences of the choice of analysis

We now go through the original conclusions drawn from the chick study and see how they could have differed, had they been based on the bottom row of Figure 1 rather than the more noisy, statistically inefficient summaries shown in the top row of that figure.

Perhaps most importantly, the overall impression of the data would have changed. Blackman et al. (1988) started off by declaring: "These results demonstrate that certain frequencies are effective ($P < .05$) in causing enhance calcium-ion efflux while others are not." And, indeed, upper-left plot of Figure 1 shows a mix of positive and negative results, and most are not statistically significant. In contrast, in the lower-left plot all the point estimates are positive, making it clear that the results are consistent with a general pattern of positive effects with uncertainty at individual frequencies.

Removing the sham correction affects more detailed conclusions as well. Blackman et al. (1988) pull out patterns from the top-right graph Figure 1 that do not appear when this same p-value classification is used in the more bottom-right graph. They label one set of responses as occurring at five frequencies at the low end—15, 45, 75, 105, and 135 Hz—but in the new graph the frequencies of 30 and 60 Hz also fall in this $p < 0.01$ category, destroying the alternating pattern of positive and null results. Relatedly, they place 60, 90, and 180 Hz in together in a set of intermediate p-values—but in the cleaner summary, this category contains 120 Hz rather than 60 Hz, obviating a discussion later in the paper of how "the data at 180 Hz could be the fundamental of a nonlinear mechanism . . . leading to subharmonic frequencies that manifest at 90 and 60 Hz."

The article also includes speculation about what is going on at 405 Hz, which in the original analysis is the only frequency at the high end with a statistically significant effect; see the top-left graph of Figure 1. The revised, bottom-left, plot tells a completely different story: the estimate at

Parameter	Estimate (s.e.)	95% interval
μ^θ	0.097 (0.015)	$[0.069, 0.126]$
σ^θ	0.069 (0.014)	$[0.044, 0.099]$
μ^b	0.004 (0.006)	$[-0.008, 0.017]$
σ^b	0.008 (0.006)	$[0.000, 0.021]$

Table 2: *Posterior means, standard deviations, and 95% intervals for the hyperparameters in the hierarchical model fit to the chick data.*

405 Hz is no longer statistically significant, but those at 420 and 450 Hz are. An entirely new set of theories would be needed to explain this pattern.

We are not saying that it was a bad idea for the authors of the original paper to engage in data-based scientific speculation. Rather, our point is that the statistically inefficient decision to adjust for the sham data is not merely of theoretical interest; it has real effects on the empirical conclusions from this study and also on the scientific explanations proposed for further study. The analysis subtracting the sham estimates may have seemed at the time like a safe choice, but in this example it simply added noise.

For an example of the practical impact of not fully modeling variation, consider this quote from Blakeslee (1991): "This requirement for exact field geometries may help explain the 'Cheshire cat phenomenon' in bioelectromagnetic experiments, Dr. Blackman said. Researchers have long been vexed by a now-you-see-it, now-you-don't problem as many experiments were not reproducible from one laboratory to the next, he said." It seems that active research effort was devoted to studying experimental differences which well may have been explainable by noise. We consider this not a criticism of these particular researchers so much as a general concern with the routine use of simple statistical analyses (in this case, subtraction of the sham estimate) which lead to unnecessarily variable conclusions.

3.6. Reanalysis using the multilevel model

We now fit the multilevel model (2) to the Blackman et al. data; inferences for the hyperparameters appear in Table 2. The estimates of μ^θ and σ^θ imply a distribution of treatment effects with a clearly positive mean, along with substantial variation, implying different effects at different frequencies. But there is no evidence for any sham effects: both μ^b and σ^b are estimated to be essentially zero—even at the highest end of the uncertainty interval, a value of 0.02 would be a tiny amount of bias compared to treatment effects that are three to six times higher. The lack of evidence for any sham effects is no surprise given the preliminary analysis shown in Figure 2. Again, the point of our hierarchical model in this example is not to discover the evident lack of noticeable sham effects but rather to be part of a general approach to this sort of problem.

Figure 6a shows the posterior mean and standard deviation of the treatment effect θ_j for each experiment j. For comparison, we display in Figure 6b the raw estimates $y_{j1} \pm \sigma_{j1}$ from the exposed data. The estimates from the hierarchical model have been partially pooled toward the common mean but otherwise show a pattern similar to that of the raw data, with the largest change being the raw estimate at 255 Hz that had a very large standard error (the long error bar in Figure 6b) and was thus pulled closer to the center of the distribution.

Again, we are not surprised that our Bayesian inferences are qualitatively similar to the raw estimates from the exposed data. Recall that this whole example came up because the standard recommendation to subtract the sham data yielded unnecessarily noisy estimates. We consider it a success that hierarchical modeling gives us a general approach to arrive at a reasonable conclusion.

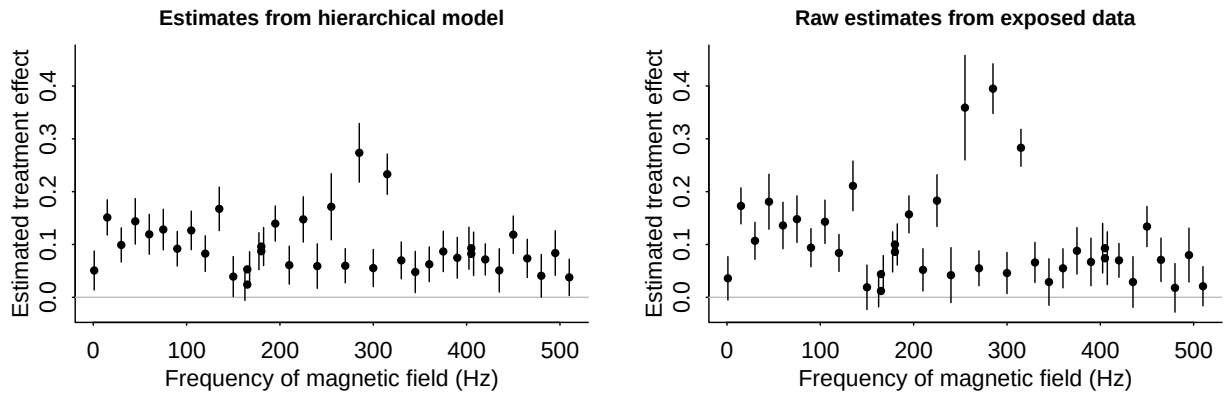

Figure 3: *(a) Posterior mean ± standard deviation of each treatment effect θ_j from the hierarchical model fit to the chick data. The fitted model estimated the sham effects to be essentially zero (see discussion of Table 2), and so these estimated treatment effects come pretty much from the exposed data alone. For three of the frequencies (165, 180, and 405 Hz) the experiment was performed twice, and in these cases we have jittered the two experimental results so they both appear on the graph. (b) For comparison, the raw estimates y_{j1} from the exposed data ± standard error. The Bayesian hierarchical estimates on the left plot are partially pooled toward a common mean.*

In the data at hand we see no clear patterns or correlations that would warrant a more structured model for the treatment effects or the biases; however, in Appendix A we consider some alternative models of the form (3), as a robustness check and also to demonstrate how such models could be fit using Stan.

4. Applied example 2: Meta-analysis of sham-controlled trials of rTMS for treating major depression

4.1. Background

Major depression is a highly prevalent public health issue with enormous social and economic cost. For a large group (20–30%) of patients, existing treatments do not suffice to achieve remission. Moreover, existing treatments can take a long time to achieve remission if they do at all and tend to be associated with unpleasant side effects. For this reason, new treatment options for major depression are badly needed. In the last two decades, repeated transcranial magnetic stimulation (rTMS) has emerged as a promising, non-invasive new treatment option. Specifically, rTMS treatment is achieved by inducing electric currents within the brain by applying a changing magnetic field (generated by electricity running through a coil of wire near the scalp of the patient).

Since its introduction as a potential treatment for depression, rTMS has been studied in a large number of randomized control trials. Recently, Berlim et al. (2014) published a highly cited systematic review and meta-analysis of such trials to assess the suitability of rTMS as a treatment for major depression, estimating the response, remission and dropout rates. This meta-analysis included 29 suitable randomized, double-blind, sham-controlled trials out of the 396 such trials they previously identified; in the present paper, we analyze the 15 of these trials that include data on remission rates. In each trial, patients were exposed to a real or sham rTMS treatment, consisting of a coil angled on the scalp or the use of a specific sham coil.

Berlim et al. (2014) report odds ratios between the real and sham treatments for response, remission, and dropout rates, for both the individual studies included in the meta-analysis as well

	Sham treatment		Real exposure	
Study name	Remission n_{j0}	Total N_{j0}	Remission n_{j1}	Total N_{j1}
George et al. (1997)	0	5	1	7
Berman et al. (2000)	0	10	1	10
⋮	⋮	⋮	⋮	⋮
Bakim et al. (in press)	1	12	9	23

Table 3: *A portion of the data used for the rTMS meta-analysis reported by Berlim et al. (2014). The data consists of remission and total counts observed in all the included studies for both real and sham rTMS treatment.*

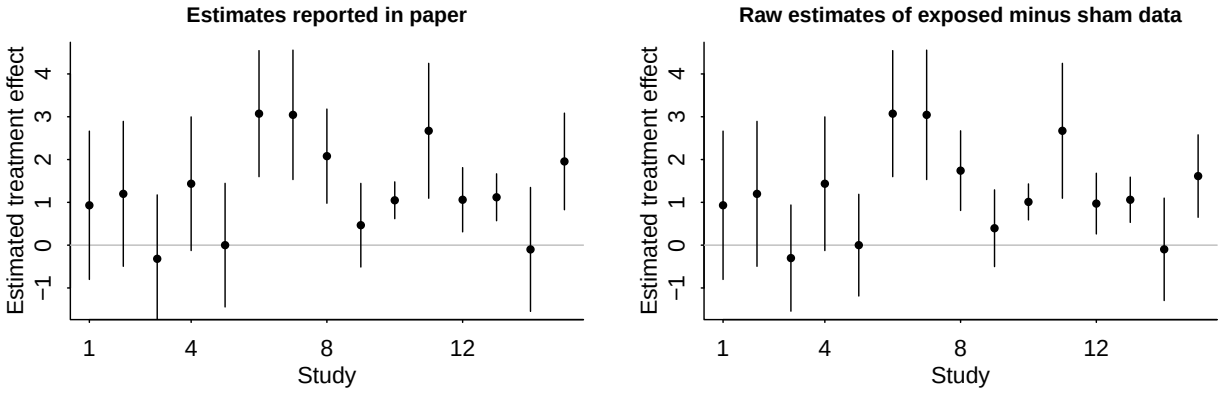

Figure 4: *(a) Estimated effects for each of the rTMS experiments as reported by Berlim et al. (2014). (b) Raw difference estimates $y_{j1} - y_{j0}$ with standard errors $(s_{j1}^2 + s_{j0}^2)^{1/2}$.*

as for the compound meta-analysis. In this paper, we will focus on an alternative analysis of the remission rates, but we could have equally well have chosen the response or dropout rates.

A subset of the data used by Berlim et al. (2014) are displayed for clarity in Table 3. For study j, it consists of remission counts n_{j0} out of a total of N_{j0} patients for the sham treatment as well as remission counts n_{j1} out of a total of N_{j1} patients for the real treatment.

4.2. Originally published analysis

From this data, Berlim et al. (2014) calculate estimates and confidence intervals of the odds ratio of the two treatments, using the hierarchical modeling approach of DerSimonian and Laird (1986). These estimates can be understood in our framework as follows.

We can straightforwardly calculate the log odds of remission $y_{j0} = \log((n_{j0} + 0.5)/(N_{j0} + 1))$ for the sham experiments and $y_{j1} = \log((n_{j1} + 0.5)/(N_{j1} + 1))$ for the real ones.[2] Assuming a binomial distribution of the remission counts, the log odds will be approximately normally distributed for large enough sample sizes. We can apply the power method to derive estimates s_{j0} and s_{j1} for the standard errors of y_{j0} and y_{j1}, respectively. To be precise, we estimate $s_{ji} = \sqrt{(n_{ji} + 0.5)^{-1} + (N_{ji} - n_{ji} + 0.5)^{-1}}$. We are now again in a position where we can think of y_{ji} arising from model (1).

Their estimates of the log odds ratio are reproduced in Figure 4a. They can be seen to be compatible with the plain difference estimates in Figure 4b. Both the estimates and standard

[2]As is commonly done, we deal with cells with zero counts by adding a Haldane-Anscombe correction of 0.5.

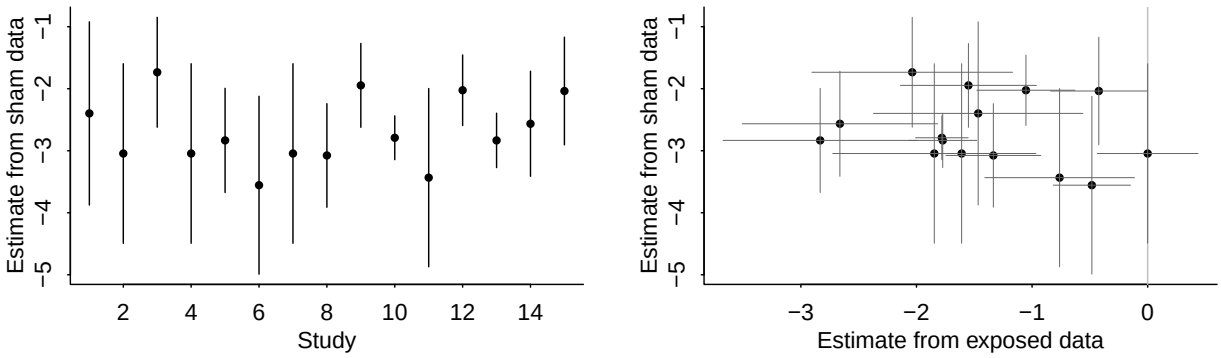

Figure 5: *(a) Estimates ± standard errors of the effect of the sham treatments for the studies in the meta-analysis of rTMS. (b) Sham vs. exposed estimates.*

Parameter	Estimate (s.e.)	95% interval
μ^θ	1.2 (0.3)	$[0.7, 1.8]$
σ^θ	0.5 (0.3)	$[0.0, 1.1]$
μ^b	-2.5 (0.2)	$[-2.9, -2.1]$
σ^b	0.3 (0.2)	$[0.0, 0.7]$

Table 4: *Posterior means, standard deviations, and 95% intervals for the hyperparameters in the hierarchical model fit to the rTMS data.*

errors corresponding to different experiments vary widely across studies (particularly considering that these are log odds).

4.3. Exploration of the sham data

The nature of the sham data in this second example is different than in the example of Section 3. Indeed, it would make not any sense for our sham data to be noise (centered at zero), as log odds of zero would correspond to even odds, that is, a coin flip. But a treatment for major depression with a remission rate of 50% would constitute a major breakthrough. Therefore, we would expect our sham data to look like anything but noise. Indeed, upon inspection, this turns out to be the case, as shown in Figure 5.

This immediately makes clear that discarding the sham data and working with the exposed data alone is not an option. However, our hierarchical model (2) still can provide a superior alternative as it allows us to reconsider our estimates of the sham and treatment effects in the individual studies in the light of the larger meta-analysis, by pooling them towards their common mean.

4.4. Reanalysis using the multilevel model

We now fit the multilevel model (2) to the Berlim et al. data; inferences for the hyperparameters appear in Table 4. The estimates of μ^θ and σ^θ imply a distribution of treatment effects with a clearly positive mean, along with substantial variation, implying different effects across different studies. We can interpret the estimated treatment effect μ^θ of 1.2 as saying that the odds of remission when receiving the real treatment are about three and a half ($\approx \exp(1.2)$) times as good as when being treated with the sham. We can interpret our estimated sham effect μ^b of -2.5 as saying that even with the sham treatment, there is still a probability of about 1/13 of remission.

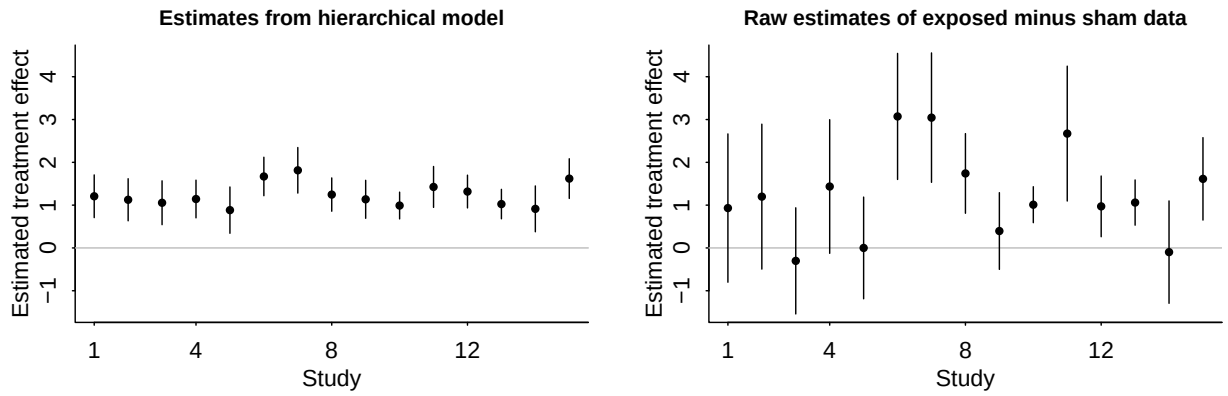

Figure 6: *(a) Posterior mean ± standard deviation of each treatment effect θ_j from the hierarchical model fit to the rTMS data. The fitted model did not estimate the sham effects to be zero, and so these estimated treatment effects take into account both the exposed and sham data. (b) For comparison, the raw difference estimates $y_{j1} - y_{j0} \pm$ standard error. The Bayesian hierarchical estimates on the left plot are partially pooled toward a common mean.*

Figure 6a shows the posterior mean and standard deviation of the treatment effect θ_j for each experiment j. For comparison, we display in Figure 6b the raw difference estimates $y_{j1} - y_{j0} \pm (s_{j1}^2 + s_{j0}^2)^{1/2}$. The estimates from the hierarchical model have been partially pooled toward the common mean but otherwise show a pattern similar to that of the difference estimates, with the largest changes being the studies that with extreme conclusions and large standard error. These were thus pulled closer to the center of the distribution.

Seeing that we fit our models in Stan, there is nothing forcing us to make a normal approximation to the likelihood and we could in fact have worked just as easily with a binomial observation model. This does not substantively alter the conclusions, though it ends up pooling the studies slightly less and leads to larger estimates of uncertainty. We discuss this in Appendix B.

5. Evaluating the competing estimates using simulation

We have seen the hierarchical model work on two real problems, one where there was no evidence of sham effects and one where correction for the sham was necessary. We can better understand how the model works by using simulation to set up and evaluate a series of scenarios with different levels of strength of the sham signal. We demonstrate this approach by perturbing the chick brain example of Section 3.

5.1. Setting up a family of hypothetical scenarios

To see what happens when different levels of sham correction is necessary, we study a series of simulated examples indexed by a parameter tied to the size of the sham effects. We can then compare the three estimates—(a) the exposed data estimate, y_{j1}, (b) the difference between exposed and sham, $y_{j1} - y_{j0}$, and (c) the hierarchical model estimate $\mathrm{E}(\theta_j|y)$—and see how they perform as a function of the scale of the bias parameters, b_j.

We set μ^b to 0 and consider a range of values for σ^b, for each performing the following steps 200 times: (1) Simulate one draw of the vector of 38 values $b_j, j = 1, \ldots, J$, drawing them independently from the normal$(0, \sigma^b)$ distribution; (2) Draw the vector of the 38 values $\theta_j, j = 1, \ldots, J$, from their

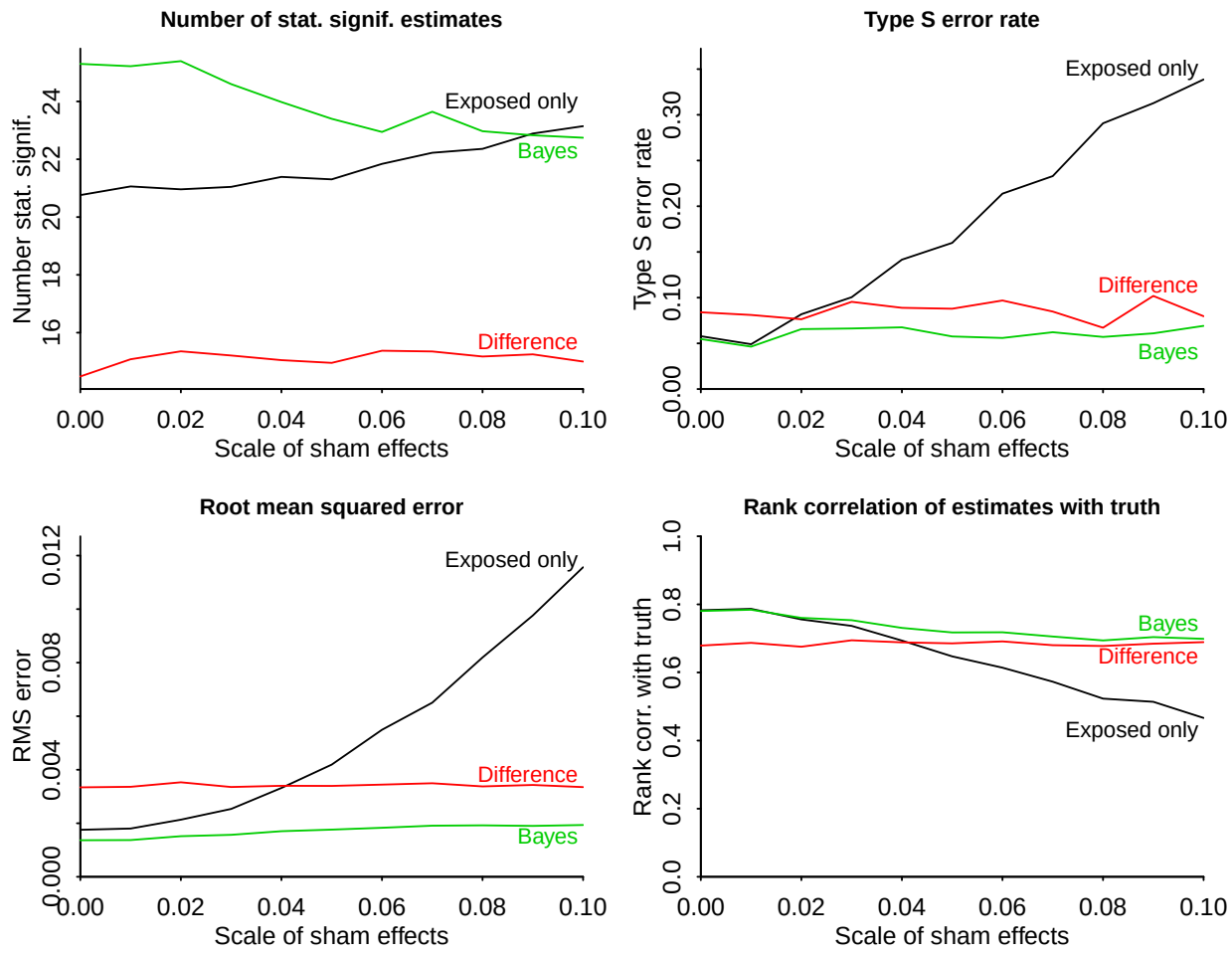

Figure 7: *Results of simulation study comparing three estimates—(a) the exposed data estimate, y_{j1}, the difference between exposed and sham, $y_{j1} - y_{j0}$, and the Bayesian hierarchical model estimate $E(\theta_j|y)$—to simulated data. The four graphs show the results for four different frequency evaluations, and on each graph the horizontal axis represents σ^b, the standard deviation of the sham effects in the simulation.*

posterior distribution from Section 3.6; (3) Simulate one dataset, that is a vector of 38 values $y_{j0} \sim \text{normal}(b_j, \sigma^y)$ and a vector of 38 values $y_{j1} \sim \text{normal}(\theta_j + b_j, \sigma^y)$.

We are assuming that the residual scales σ^y are known and all equal to 0.04, a value chosen because it is approximately the average of the standard errors in the data; see Table 1. This simplification, along with that of assuming $\mu^b = 0$, makes it easier to interpret the results of our simulation but should not materially affect our results. We explore the role of μ^b in Section 6.

5.2. Evaluations

As discussed in Section 2.3, for each set of simulated parameters and data, we then compute the following four summaries for each of the three estimates: the proportion of the 38 estimates that are statistically significant, the type S error rate, the mean squared error of the 38 estimates, and the rank correlation between the estimates and the true θ_j's.

We choose a grid of values of σ^b between 0 and 0.10, choosing that upper bound as this is the

approximate standard deviation of treatment effects (see inference for σ^θ in Table 2), and we would not expect the variation in sham effects to be higher than the variation in treatment effects.

For each σ^b we average each of the above summaries over our 200 simulations to obtain 3×4 matrix of four frequency evaluations for the three estimates: exposed, exposed minus sham, and hierarchical Bayes. Figure 7 plots the results for each frequency evaluation as a function of σ^b, the scale of the sham effects. The difference estimate, $y_{j1} - y_{j0}$, outperforms the exposed-only estimate y_{j1} when sham effects are large (the right side of each graph) but not when sham effects are small. The Bayesian hierarchical estimate outperforms both, in part by appropriately managing the sham data and in part by pooling across experiments.

We now go through each of the frequency properties for this example:

- *Number of statistically significant claims:* The difference estimate yields the lowest rate of statistically significant results, which makes sense given that it is the noisiest of the estimates. When sham effects become large, the rate of apparently statistically significant estimates from the exposed data alone goes up, but this is an illusion based on the fact that the variance in the data is increasing but this is not reflected in the standard errors.

- *Type S error rate:* The difference and Bayes estimates have approximate 5% type S error rates, as does the exposed-only estimate when the sham effects are negligible. As sham effects become larger, the error rate for the exposed-only estimate becomes increasingly unacceptable.

- *Root mean squared error:* The Bayes estimate performs the best, unsurprisingly as it makes use of the most information, and we are simulating from the model. The exposed-only estimate outperforms the difference estimate when sham effects are near zero—this is what we saw in Section 3—but when sham effects are large, the exposed-only estimate has a huge error.

- *Rank correlation with truth:* When sham effects are small, the exposed-only estimate is best; when sham effects are large, the difference is best; in all cases the Bayes performs as well as the other two. At each extreme, the Bayes does as well as, not better than the corresponding simple estimate; this is because, in this simple simulation where the error variances for all experiments are equal, the partial pooling across experiments affects estimates and standard errors but does not alter the ranking of the 38 estimates.

The results shown in Figure 7 are consistent with the idea of the difference being a conservative estimate—and, indeed, *if* the only available choices were the exposed-only and the difference estimate, *and* no information were available regarding σ^b, the scale of the sham effects, then we might well prefer the difference as the safe option. In fact, though, we are also free to use the hierarchical Bayes estimate, and even if that were not available, the data are informative about σ^b, so we would not recommend the difference estimate as a default analysis.

6. Linear adjustment via partial Bayesian inference

We can gain intuition about the sham-adjustment problem by fitting a partially Bayesian model in which the sham effects come from a normal(μ^b, σ^b) distribution but the treatment effects θ_j are estimated using maximum likelihood (equivalently, Bayesian inference with σ^θ set to infinity and μ^θ becoming irrelevant). This can also be viewed as a measurement error model, where the y_{j0}'s are noisy measurements of latent variables b_j. To simplify the algebra, we assume a normal likelihood for the measurements.

Under any of these formulations, fixing the hyperparameters results in linear estimates for the θ_j's, which in turn allows clear comparisons with the exposed-only and difference estimates. This

is related to the work of Turner et al. (2015) to make Bayesian meta-analysis more accessible using analytic formulas that approximate fully Bayesian inferences.

To work out the solution algebraically it is convenient to first perform inference for the sham effects. Combining the prior distribution, $b_j \sim \text{normal}(\mu^b, \sigma^b)$, with the sham measurement, $y_{j0} \sim \text{normal}(b_j, \sigma_{j0})$, yields a posterior distribution, $b_j \sim \text{normal}(\hat{b}_j, s_j)$, where

$$\hat{b}_j = \frac{\frac{1}{(\sigma^b)^2}\mu^b + \frac{1}{\sigma_{j0}^2}y_{j0}}{\frac{1}{(\sigma^b)^2} + \frac{1}{\sigma_{j0}^2}} \qquad \text{and} \qquad s_j = \left(\frac{1}{(\sigma^b)^2} + \frac{1}{\sigma_{j0}^2}\right)^{-1/2}.$$

The corresponding maximum likelihood estimate $\hat{\theta}_j$ is $y_{j1} - \hat{b}_j$, which can be written as

$$\hat{\theta}_j = y_{j1} - \mu^b - \lambda(y_{j0} - \mu^b) \tag{4}$$

with standard error $\sqrt{s_j^2 + \sigma_{j1}^2}$, and where

$$\lambda = \frac{(\sigma^b)^2}{(\sigma^b)^2 + \sigma_{j0}^2}$$

is the variance ratio which determines the amount by which the exposed-data estimate must be adjusted for the sham measurement.

The estimate (4) reduces to the exposed-only estimate when $\mu^b = \sigma^b = 0$ (that is, when there are no sham effects) and reduces to the difference estimate as $\sigma^b \to \infty$ (as sham effects become large). The standard error of $\hat{\theta}_j$ reduces to σ_{j1} when sham effects are zero and $\sqrt{\sigma_{j0}^2 + \sigma_{j1}^2}$ in the limit of large sham effects.

In between these extremes, equation (4)—the maximum likelihood estimate under the measurement error model—is constructed by first subtracting the average sham effect, which represents the average bias for all the experiments—and then a subtracting a fraction of the relative estimated sham effect from experiment j, with that fraction depending on the relative values of σ^b and σ_{j0}. For the chicken data, μ^b is estimated to be essentially zero and σ^b is estimated to be much smaller than σ_{j0} for all the experiments (see Table 2) , so there is no essentially no need to adjust for the sham measurements.

In practice we would recommend full Bayesian inference as in Section 2.2. Or, if there is reluctance to partially pool across experiments, one could fit the same Bayesian model but removing the prior on the θ_j's (equivalently, constraining σ^θ to ∞). The point of the above algebra is just to clarify the way in which the optimal estimate of treatment effects will in general approximately take the observed estimate y_{j1} and subtract some fraction, between 0 and 1, of the sham estimate.

7. Discussion

7.1. Failure modes and limitations of the method

For the reasons discussed above, we prefer the hierarchical Bayesian model to the alternative analyses for the chick brain study: we think that the estimates obtained from our model are more reasonable and that they would yield better predictions in a replication study. But there must be settings where our approach would perform poorly. When will that occur?

Speaking generally, Bayesian inference with noisy data works by partial pooling toward a fitted model. When the fitted model is wrong, the pooling can go in the wrong direction, yielding poor

inferences. In the problem discussed in this paper, the sham and treatment estimates are each pooled toward the mean of that set of experiments. For the sham, this does not seem to be a problem, first because we expect sham effects to be small, second because we have no reason to expect patterns in the sham effects. If we did expect such patterns, it would make sense to include them in the model, for example by allowing a correlation between sham and treatment effects as discussed in Appendix A.1. For the treatment effects, partial pooling toward a common mean could be more of a concern, for example if there is a trend or if the pattern of effects is otherwise predictable. This is related to the problem of edge effects when estimating a function from noisy data: an extrapolative model can overfit trends in the data, but a model that is more conservative in its extrapolation can flatten out at the edges. Ultimately one must accept that inferences are sensitive to uncheckable assumptions.

For our default hierarchical model to fail badly, two things must happen: the data must be noisy enough for the partial pooling to make a difference, and the underlying trend or pattern must itself be strong. Both these things can happen, for example if the treatment effects follow a linear trend. In our two applied examples, there was no apparent trend in the data; had there been, it surely would have been included in the model. But a proposed statistical method will be used in all sorts of settings. Were we to fit our no-trend hierarchical model to data with an actual trend, we would overestimate effects at the low end and underestimate at the high end, in aggregate understating the variation in the treatment effects. In this case, our recommended solution would be to incorporate this possible trend by adding it into the mean of the distribution for θ_j in (2).

More generally, nonlinear models are possible, hence it can make sense to check sensitivity of analyses to various choices of model, as we demonstrate in Appendix A. We prefer our default hierarchical model to the simple default of exposed minus sham, but in general it makes sense to consider scientifically plausible alternatives as well. This is an unavoidable concern when using measurement error or latent variable models, but ultimately we see no good alternative to modeling, as the simple unpooled estimates are just too noisy and wasteful of data.

The main practical limitation of our method is that it works best when there is a large number of repeated studies. When the number of studies J is small, the hierarchical model can still be fit, but the user would be advised to include strong prior information on the hyperparameters. This could well be a good idea—indeed, we would prefer it to a riskier or noisier strategy such as fully subtracting or ignoring the sham data—but we recognize that a fully Bayesian approach would put more of a burden on many researchers. Hence in the present paper we focus our recommendations on the problem of repeated studies.

7.2. Chick brains experiment

A key point of this paper is that our analysis could have made a difference in our motivating example.

Blackman (2015) wrote that his team "worked very closely with a statistician ... to optimize our procedures for maximum statistical power." Care went into both the scientific and statistical aspects of the design of the study, as well as the data collection itself. This is one indication of the potential importance of the statistical modeling and analysis plan we have presented here: if a team of conscientious researchers, working on a policy-relevant research program and aware of cost constraints and the importance of statistical efficiency, can perform an analysis that is mathematically equivalent to discarding half the information in their data, this represents large gains from a new paradigm, moving away from cookbook rules to an open-ended modeling approach. Indeed, Blackman (2005) also writes, "Plans were made to follow up ... but the experiment could not be brought to fruition." In this case discarding the sham data would have been equivalent to

doubling the sample size of the experiments, without any data-collection cost at all.

In retrospect it would have been enough to collect a smaller set of sham data in the chick brains study; there was no need to replicate all 38 experiments. This was not clear a priori but is apparent upon examination of the sham data. A more efficient, sequential, design would recommend gathering *some* sham data, but once its irrelevance becomes clear (as seen from the estimated values of μ^b and σ^b in the hierarchical model), not so much would need to be collected. Given the uncertainty in some of the frequency comparisons of interest (an uncertainty masked by the illusion of informativeness of comparisons of p-values), limited experimental resources could have been used more effectively by collecting more data on non-sham treatments. We do not consider this as a devastating criticism of the study—it is unfortunately all too common, including in our own work, to gather data in rectangular structure with an eye toward convenience rather than efficiency—but it is worth considering these issues when designing future experiments.

7.3. More general implications for design and analysis of structured experiments

What is striking about the results from this paper, as distinguished from many other examples of the practical efficiency gains that can be obtained from Bayesian inference, is how simple and effectively the Bayesian approach works out in this example, requiring no specialized knowledge or custom prior distributions. This gives us hope that hierarchical modeling can resolve other common data-combination problems in applied statistics, and it is why we have been continuing to chew on this example for thirty years.

Specifically, we recommend our Bayesian multilevel model as a default analysis for repeated controlled experiments. Indeed, it gives more efficient estimates than both the commonly used difference or exposed-only estimates. More importantly still, it systematically determines from the data how much adjustment for the sham measurements is appropriate, by interpolating between the extremes of difference and exposed-only estimates, rather than leaving that choice to the scientist.

There is some awkwardness that, in order to perform our more efficient estimate, we need to model the treatment effects θ_j and the biases b_j. This is a general property of measurement-error models, and we believe there are many cases, including the examples described in the paper, where the small effort in constructing probability models for treatment effects and biases is minor compared to the efficiency gains obtainable from the model-based estimate. An alternative for those who would prefer not to partially pool the θ_j's across experiments is to use the procedure of Section 6 to find an optimal linear adjustment, which is corresponds to modeling just the biases without partially pooling the treatment effects.

On top of that, we suggest a sequential experimental design, in case of costly sham data collection. If, in the course of data collection, the recommended analysis confidently estimates the sham effects to be substantively insignificant, collection of sham data can be halted and the resources can be transferred, for instance, to collection of more exposed data.

Combining these two recommendations for the statistical analysis and experimental design of controlled experiments should enable a more cost-effective scientific practice. We hope this will contribute towards in an increase in replicable scientific findings.

References

Ashenfelter, O., Zimmerman, P., and Levine, D. (2003). *Statistics and Econometrics: Methods and Applications*. New York: Wiley.

Berlim, M. T., et al. (2014). Response, remission and drop-out rates following high-frequency repetitive transcranial magnetic stimulation (rTMS) for treating major depression: A systematic

review and meta-analysis of randomized, double-blind and sham-controlled trials. *Psychological Medicine* **44**, 225–239.

Blackman, C. (2015). Replication and extension of Adey group's calcium efflux results. In *Electromagnetic Fields in Biology in Medicine*, ed. M. S. Markov, 7–14. Boca Raton, Fla.: CRC Press.

Blackman, C. F., Benane, S. G., Elliott, D. J., House, D. E., and Pollock, M. M. (1988). Influence of electromagnetic fields on the efflux of calcium ions from brain tissue in vitro: A three-model analysis consistent with the frequency response up to 510 Hz. *Bioelectromagnetics* **9**, 215–227.

Blakeslee, S. (1991). Electromagnetic fields are being scrutinized for linkage to cancer. *New York Times*, 2 Apr, C3.

Brodeur, P. (1989a). The hazards of electromagnetic fields I—Power lines. *New Yorker*, 12 June, 51–88.

Brodeur, P. (1989b). The hazards of electromagnetic fields II—Something is happening. *New Yorker*, 11 June, 47–73.

Brodeur, P. (2000). *Currents of Death*. New York: Simon and Schuster.

DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

Fuchikami, M., et al. (2010). Epigenetic regulation of BDNF gene in response to stress. *Psychiatry Investigation* **7**, 251.

Gelman, A., and Stern, H. S. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician* **60**, 328–331.

Greenland, S. (2005). Multiple bias modelling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society A* **168**, 267–306.

Higgins, J. P. T., and Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15**, 2733–2749.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240.

Kádár, E., Lim, L. W., Carreras, G., Genís, D., Temel, Y., and Huguet, G. (2011). High-frequency stimulation of the ventrolateral thalamus regulates gene expression in hippocampus, motor cortex and caudate-putamen. *Brain Research* **1391**, 1–13.

Lehmann, E. L., and Scheffe, H. (1950). Completeness, similar regions, and unbiased estimation. I. *Sankhya* **10**, 305–340.

Le Quément, C., et al. (2012). Whole-genome expression analysis in primary human keratinocyte cell cultures exposed to 60 GHz radiation. *Bioelectromagnetics* **33**, 147–158.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* **14**, 2685–2699.

Stan Development Team (2012). Stan: A C++ library for probability and sampling. `http://mc-stan.org/`

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., and Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine* **34**, 984–998.

Wager, S., and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**, 1228–1242.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *American Statistician* **70**, 129–133.

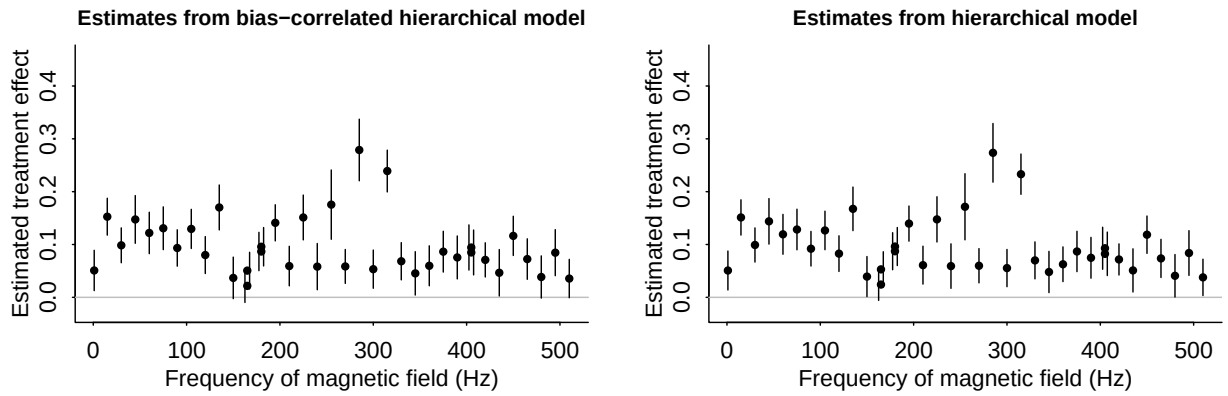

Figure 8: *(a) Posterior mean ± standard deviation of each treatment effect θ_j from the hierarchical model with correlated bias and treatment effects fit to the chick data. (b) For comparison, the estimates of the original hierarchical model without correlations, which can be seen to be almost identical.*

A. Alternative models for example 1

In this appendix, we discuss some alternative models we could have used for analyzing the data of example 1 in this paper.

A.1. Measurement error with correlation

It is conventional with measurement error models to use independent errors, and this is what we did in (2), with the idea being that there can be an average sham effect and variation in the sham effects, but with no correlation expected with the treatment effects. This makes sense in the chick experiment, as the treatment effect varies by frequency of the magnetic field, whereas the bias or sham effect should have nothing to do with frequency.

More generally, though, one might want to allow the treatment effect and its measurement bias to be correlated, in which case (2) can be generalized to a bivariate normal distribution for (θ_j, b_j) with a covariance matrix. Figure 8 shows the results of fitting this to the chick data; these treatment effect estimates are essentially the same as from the uncorrelated-errors model fit in Section 3.6.

A.2. Gaussian process for the treatment effects

A potential concern regarding the models fit so far is that they do not encode any structure in the treatment effects. One challenge here is that so many different structures are possible, as discussed in the original Blackman et al. (1988) paper. As discussed in Section 3.2, various complicated patterns of alternating frequencies were extracted, but many of these conclusions are shaky as they rely on inherently noisy comparisons of p-values.

We think the measurement error model of Section 2.2 is a sensible default analysis, but more structured models would be possible. As an example, we consider two Gaussian process (GP) models for the vector θ as a function of frequency: one model which favors local smoothness of the treatment effects (a GP with a squared-exponential covariance function) and one which favors similar effects for frequencies separated by 30 Hz (a GP with a periodic covariance function with a period of around 30 Hz).

The resulting estimates are displayed in Figure 9. The squared-exponential GP model gives

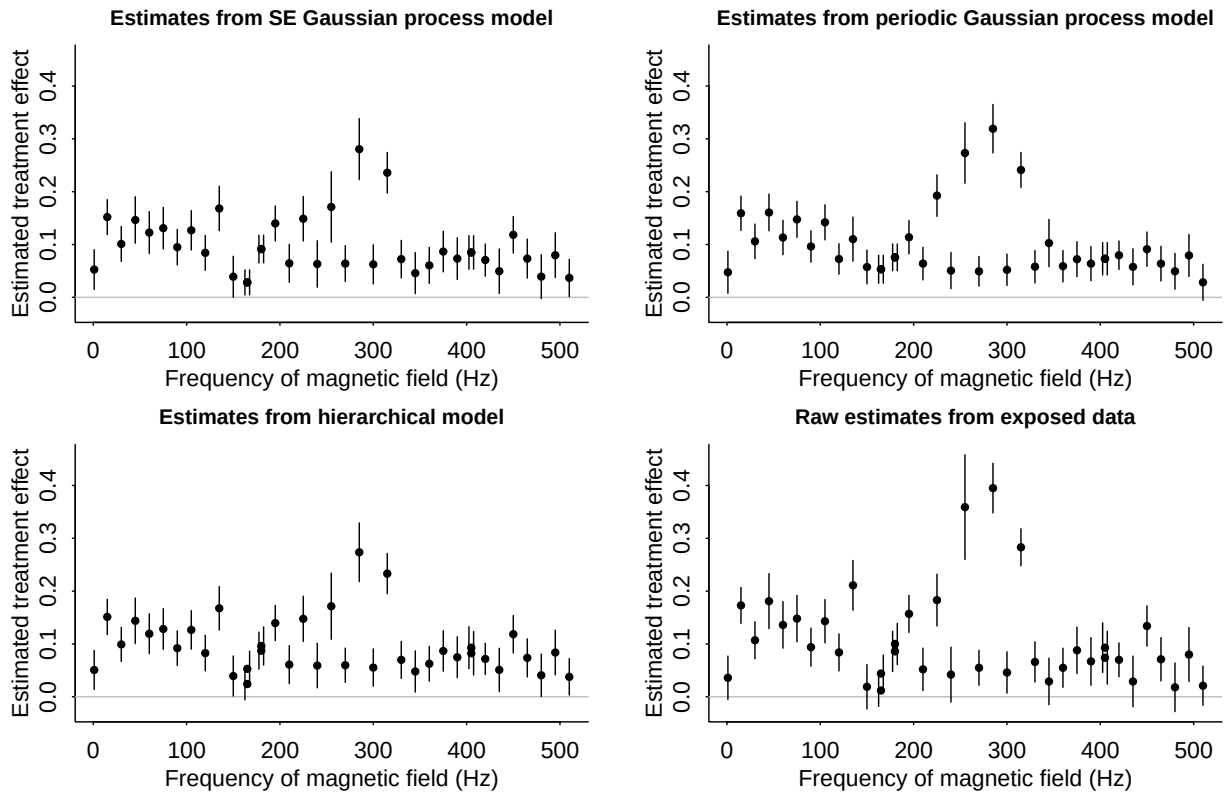

Figure 9: Top row: *(a) Posterior mean ± standard deviation of each treatment effect θ_j from the squared-exponential kernel Gaussian process model fit to the chick data. (b) The corresponding plot for the periodic kernel Gaussian process model. The fitted models estimate the sham effects to be essentially zero, and so these estimated treatment effects come pretty much from the exposed data alone. The Bayesian estimates in the left plot are partially pooled towards each other for close frequencies. The estimates in the right plot are partially pooled towards each other for frequencies whose difference is close to 30 Hz.*
Bottom row: *For comparison, (c) the estimates from our default analysis and (d) the raw estimates y_{j1} from the exposed data.*

estimates that are very close to those of the hierarchical model. This model favors stronger pooling between measurements which are close in frequency. This results, for example, in a slightly higher estimate for the treatment effect at 285 Hz but it is most visible for the frequencies with repeated measurements which it forces to have the same estimated effect sizes. For the periodic GP model, we observe interestingly different estimates compared to our default analysis, due to the periodic partial pooling behaviour it enforces. For example, we see that the estimates at 225 and 345 Hz are pulled upwards, a phenomenon we do not observe in our default analysis.

One difficulty in using such GP models for analyzing the data is the question of how to choose an appropriate prior on the length-scale parameter. This parameter regulates the scale on which the smoothing happens. That is to say, it determines how close two frequencies need to be to each other in order to qualify to be pooled together. This prior should be chosen based on domain expertise in each particular application. We believe this makes the GP analyses less suitable as a default choice, unless strong domain knowledge of that kind is available.

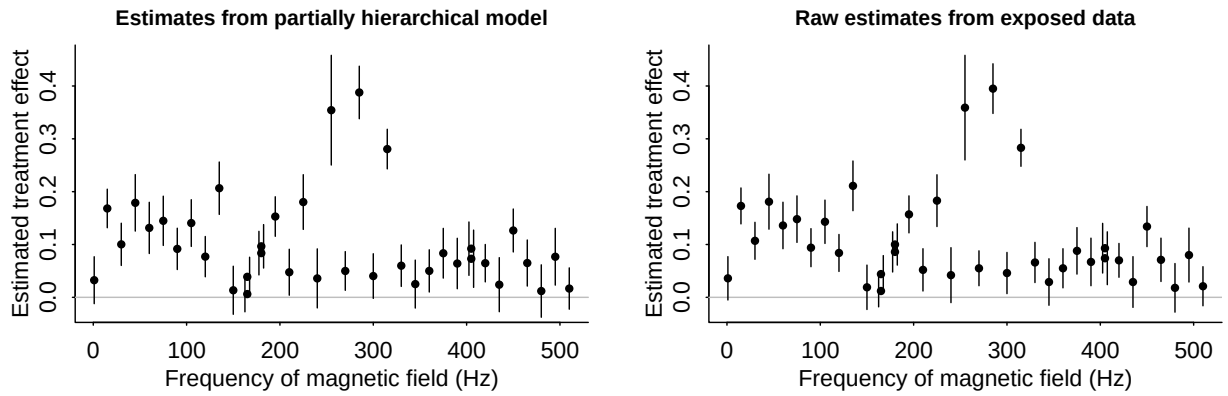

Figure 10: *(a) Posterior mean ± standard deviation of each treatment effect θ_j from the partially-hierarchical model, which partially pools the biases but not the treatment effects, as fit to the chick data. The fitted model estimated the sham effects to be essentially zero (see discussion of Table 2), and so these estimated treatment effects come pretty much from the exposed data alone. (b) For comparison, the raw estimates y_{j1} from the exposed data. These two estimates roughly coincide.*

A.3. Removing partial pooling

Following Section 6, it may be interesting to inspect the estimates given by variants of the hierarchical model, where we first remove the partial pooling of the treatment effects and next also that of the biases.

When we remove the partial pooling of the treatment effects (equivalent to the limit, $\sigma^\mu \to \infty$), but keep partial pooling of the biases, we estimate that μ^b and σ^b are both near zero, as would be expected from Table 1. As anticipated by the algebra of Section 6, we obtain, in effect, the raw exposed-only estimates. This is shown in Figure 10.

When we additionally remove the partial pooling of the biases (equivalent to the limit, $\sigma^b \to \infty$), the algebra of Section 6 would predict that we roughly end up giving the raw difference estimate. Indeed, we see this confirmed in Figure 11. The difference estimate is similar to the exposed-only estimate but is much higher in uncertainty. Dropping the partial pooling of the biases has the same result of increasing the noise in our estimates.

The two estimates of Figure 11 always coincide, but the collapse of the two estimates of Figure 10 only happens when the sham data is effectively noise. The partially hierarchical model that partially pools the biases but not the treatment effects might be a superior alternative to the exposed-only and difference estimates in case there is reluctance to partially pool across experiments as we do in our default analysis.

A.4. Alternative simulation study based on raw estimates

One objection the reader might have to our simulation study of Section 5.1 is that we were simulating the θ_j from the posterior fit from the Bayesian model, which might give the Bayesian estimates an unfair advantage. In this section, to address this concern, we show that we observe the same phenomena as discussed in Section 5.1 even if we use the raw estimates for θ_j instead.

Specifically, we perform this alternative simulation exactly as before described except that we perform the following two steps instead of steps 1–3 in Section 5.1: (1) Simulate one draw of the vector of 38 values $b_j, j = 1, \ldots, J$, drawing them independently from the normal$(0, \sigma^b)$ distribution; (2) Simulate one dataset, that is a vector of 38 values $y_{j0} \sim t_{n_{j0}-1}(b_j, \sigma^y)$ and a vector of 38 values

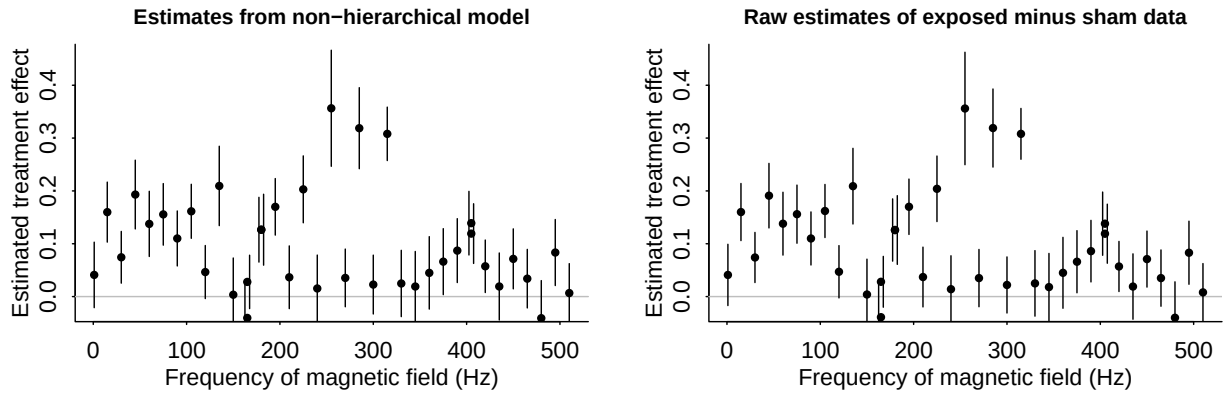

Figure 11: *(a) Posterior mean ± standard deviation of each treatment effect θ_j from the non-hierarchical model, which partially pools neither the biases nor the treatment effects, as fit to the chick data. (b) For comparison, the raw difference estimate $y_{j1} - y_{j0}$ for the chick data ± standard error. These two estimates roughly coincide.*

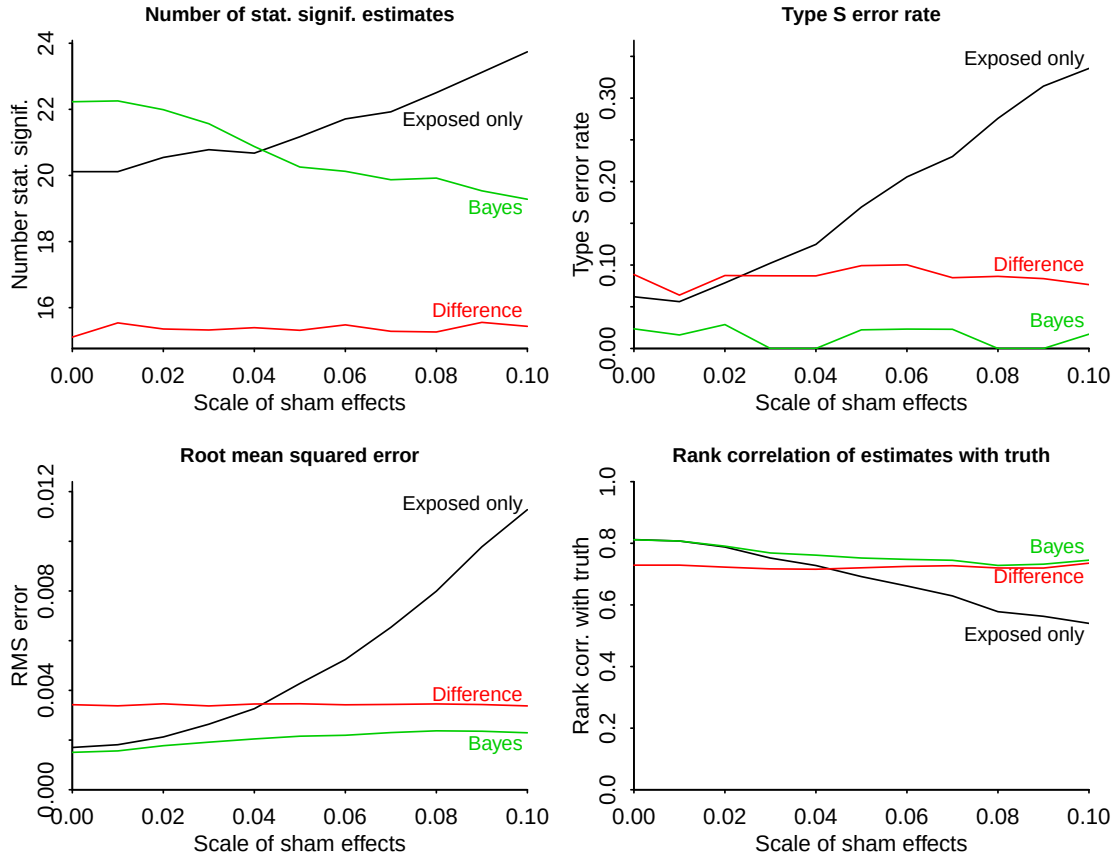

Figure 12: *Results of the alternative simulation study from Appendix A.4 comparing three estimates—(a) the exposed data estimate, y_{j1}, the difference between exposed and sham, $y_{j1} - y_{j0}$, and the Bayesian hierarchical model estimate $E(\theta_j|y)$—to simulated data. The four graphs show the results for four different frequency evaluations, and on each graph the horizontal axis represents σ^b, the standard deviation of the sham effects in the simulation.*

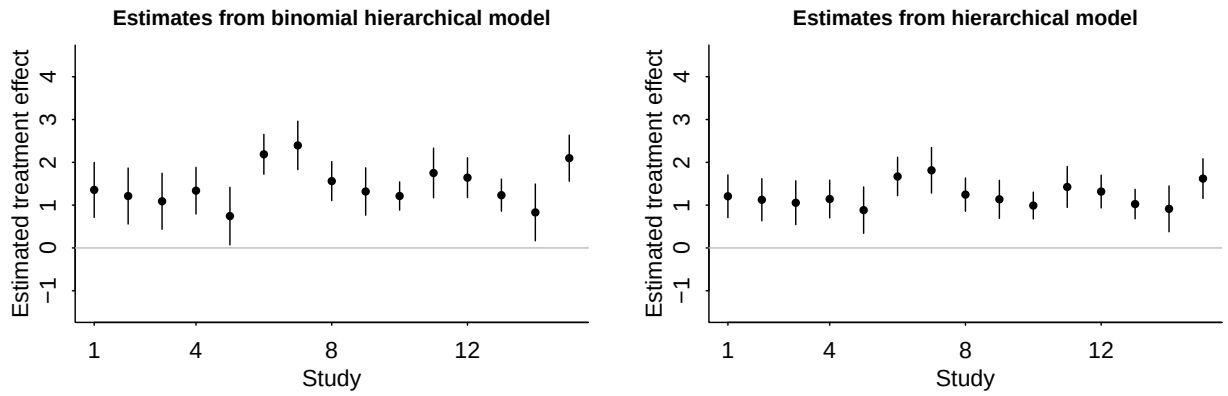

Figure 13: *(a) Posterior mean ± standard deviation of each treatment effect θ_j from the hierarchical model 5 with binomial likelihood. (b) For comparison, the same estimates from the hierarchical model 2 with normal likelihood.*

$y_{j1} \sim t_{n_{j1}-1}(y_{1j}^{obs} + b_j, \sigma^y)$, where we write y_{j1}^{obs} for the raw exposed-only estimates of the treatment effects from the actual chick data as observed by Blackman et al. (1988). We are thus centering our estimated treatment effects at the observed data rather than, as before, at the hierarchical Bayes estimates.

The results are summarized in Figure 12. They tell mostly the same story as we saw in our original simulation study. One difference is that the exposed-only estimate now consistently results in more statistically significant estimates compared to the Bayesian estimate. However, inspection of the type S error rates reveals that these extra significant estimates are not to be trusted. The results of this simulation show that the hierarchical Bayesian estimate is still superior to its two alternatives even in cases where its partial partial pooling behaviour is not an advantage.

B. Alternative model for example 2

In this appendix, we discuss an alternative model we could have used for analyzing the data of example 2. In Section 4 we used the hierarchical normal model using a standard correction for zero counts. But, given that we fit our models in Stan, we could just have easily have modeled the discrete data more directly, using a binomial likelihood rather than a normal approximation. This leads to the following model, directly modeling n_{ji}, rather than y_{ji}:

$$\begin{aligned} b_j &\sim \text{normal}(\mu^b, \sigma^b) \\ \theta_j &\sim \text{normal}(\mu^\theta, \sigma^\theta) \\ n_{j0} &\sim \text{binomial}(N_{j0}, \text{logit}^{-1}(b_j)) \\ n_{j1} &\sim \text{binomial}(N_{j1}, \text{logit}^{-1}(\theta_j + b_j)). \end{aligned} \tag{5}$$

Figure 13 shows how this leads to largely the same conclusions, but slightly less pooling and higher uncertainty than the hierarchical model with the normal likelihood function.

We summarize the estimated hyperparameters in Table 5. There are slight differences with the estimates of Table 4. Indeed, the estimates of μ^θ and σ^θ have increased while those for μ^b and σ^b have decreased in the binomial model. In particular, this results in lower odds of remission for patients receiving the sham treatment as well as a larger relative effect of the real treatment

Parameter	Estimate (s.e.)	95% interval
μ^θ	1.5 (0.3)	$[0.8, 2.2]$
σ^θ	0.7 (0.3)	$[0.2, 1.4]$
μ^b	-3.0 (0.2)	$[-3.5, -2.5]$
σ^b	0.3 (0.2)	$[0.0, 0.7]$

Table 5: *Posterior means, standard deviations, and 95% intervals for the hyperparameters in the binomial hierarchical model fit to the rTMS data.*

compared to the sham treatment. Moreover, the binomial model reports larger standard errors for all hyperparameters.