# Why we (usually) don't have to worry about multiple comparisons*

Andrew Gelman† Jennifer Hill‡ Masanao Yajima§

April 21, 2008

**Abstract**

The problem of multiple comparisons can disappear when viewed from a Bayesian perspective. We propose building multilevel models in the settings where multiple comparisons arise. These address the multiple comparisons problem and also yield more efficient estimates, especially in settings with low group-level variation, which is where multiple comparisons are a particular concern.

Multilevel models perform partial pooling (shifting estimates toward each other), whereas classical procedures typically keep the centers of intervals stationary, adjusting for multiple comparisons by making the intervals wider (or, equivalently, adjusting the $p$-values corresponding to intervals of fixed width). Multilevel estimates make comparisons more conservative, in the sense that intervals for comparisons are more likely to include zero; as a result, those comparisons that are made with confidence are more likely to be valid.

Keywords: Bayesian inference, hierarchical modeling, multiple comparisons, type S error, statistical significance

## 1 Introduction

Researchers from nearly every social and physical science discipline have found themselves in the position of simultaneously evaluating many questions, testing many hypothesis, or comparing many point estimates. In the program evaluation world this arises, for instance, when comparing the impact of several different policy interventions, comparing the status of social indicators (test scores, poverty rates, teen pregnancy rates) across multiple schools, states, or countries, examining whether treatment effects vary meaningfully across different subgroups of the population, or examining the impact of a program on many different outcomes.

*We thank the participants at the NCEE/IES multiple comparisons workshop for helpful comments and the National Science Foundation, National Institutes of Health, and Columbia University Applied Statistics Center for financial support.

†Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, www.stat.columbia.edu/∼gelman

‡Department of International and Public Affairs, Columbia University, New York, jh1030@columbia.edu, www.columbia.edu/∼jh1030

§Department of Statistics, Columbia University, New York, yajima@stat.columbia.edu, www.stat.columbia.edu/∼yajima

The main multiple comparisons problem is that the probability a researcher wrongly concludes that there is at least one statistically significant effect across a set of tests, even when in fact there is nothing going on, increases with each additional test. This can be a serious concern in classical inference and many strategies have been proposed to address the issue (see Hsu, 1996, or Westfall and Young, 1993, for reviews). A related multiple comparisons concern is that, in a setting where nonzero true effects do exist, a researcher applying multiple tests may identify additional statistically significant effects that are not in fact real.

Our approach as described in this paper has two key differences from the classical perspective. First, we are typically not terribly concerned with Type 1 error because we rarely believe that it is possible for the null hypothesis to be strictly true. Second, we believe that the "problem" is rather that we haven't properly accounted for the relationship between the corresponding parameters of the model. Once we work within a Bayesian multilevel modeling framework and model these phenomena appropriately, we are actually able to get more reliable point estimates. A multilevel model shifts point estimates and their corresponding intervals toward each other (by a process often referred to as "shrinkage" or "partial pooling"), whereas classical procedures typically keep the point estimates stationary, adjusting for multiple comparisons by making the intervals wider (or, equivalently, adjusting the $p$-values corresponding to intervals of fixed width). In this way, multilevel estimates make comparisons appropriately more conservative, in the sense that intervals for comparisons are more likely to include zero. As a result we can say with confidence that those comparisons made with multilevel estimates are more likely to be valid. At the same time this "adjustment" doesn't sap our power to detect true differences as many traditional methods do.

Rather than correcting for the problems that can arise when examining many comparisons (performing many significance tests), when we work within the Bayesian paradigm all of the relevant research questions can be represented as parameters in one coherent multilevel model. Simply put, rather than correcting for a perceived problem, we just build the right model from the start (for similar perspectives see Louis, 1984, and Greenland and Robins, 1991) This puts more of a burden on the model, and a key goal of this paper is to demonstrate the effectiveness of our suggestion for realistic examples.

Sections 2 and 3 present the multiple comparisons problem from the classical and Bayesian perspectives, respectively. Both are described within the context of a example and then potential solutions are outlined. We bolster our argument against traditional multiple comparisons corrections in Section 4 of this article through a series of small examples that illustrate several of the scenarios described above. Section 5 concludes.

# 2 Multiple comparisons problem from a classical perspective

## 2.1 Infant health and development study

In this section we walk through a relatively simple example using data from a real study to illustrate the issues involved in performing multiple comparisons from both a classical perspective and a Bayesian multilevel model perspective. We're using data from the Infant Health and Development Program, an intervention that targeted children who were born premature and with low birth weight and provided them with services such as home visits and intensive high quality child care. The program was evaluated using an experiment in which randomization took place within site and birth weight group. The experimental design was actually slightly more complicated (as described in Infant Health and Development Program, 1990) but we're going to keep things simple for expository purposes. In fact, for this first illustration we will assume that it was a simple randomized block experiment with the eight sites as blocks.

In this context, we're not just interested in the overall treatment effect. Given that the composition of participating children was quite different across sites and that program implementation varied across sites as well, we would like to know for *each site individually* whether or not a statistically significant effect was present. However, we may be concerned that, in the process of conducting eight different significance tests, we are misperceiving our overall risk of making a false claim. This overall risk of error is sometimes referred to as the familywise error rate (Tukey, 1953). A similar problem arises if we are interested in comparing whether there are significant differences in treatment effects across sites.

## 2.2 Classical perspective

A classical model fit to these data might look like this

$$\begin{aligned} y_i &= \sum_j (\gamma_j S_i^j + \delta_j S_i^j P_i) + \epsilon_i, \\ \epsilon_i &\sim \mathrm{N}(0, \sigma^2), \end{aligned}$$

where $y_i$ denotes student $i$'s test score, $S_i^j$ in an indicator for living in site $j$, and $P_i$ is an indicator for program status. Although this may not be the most common way to specify this model, it is useful because here $\delta_j$ represents the treatment effect in the $j^{\text{th}}$ site and $\gamma_j$ represents the average test score for the untreated in each site.[1] This allows us to directly test the significance of each site effect.

[1] Birthweight group was also added as a predictor in this model but we ignore it in this description for simplicity of exposition.

For any given test of a null hypothesis, say $H_0^j : \delta_j = 0$, versus an alternative, say $H_A^j : \delta_j \neq 0$, there is a 5% chance of incorrectly rejecting $H_0^j$ when in fact it is true. If we test two independent hypotheses at the same significance level ($\alpha = .05$) then the probability that at least one of these tests yields an erroneous rejection raises to $1 -$ Pr(neither test yields an erroneous rejection of the null) $= 1 - .95 * .95 = .098 \approx .10$. Following the same logic, if we performed (independent) tests for all 8 sites at a .05 significance level there would be a 34% chance that at least one of these would reject in error.

### 2.3 Bonferroni correction

One of the most basic and historically most popular fixes to this problem is the Bonferroni correction, which adjusts the $p$-value at which a test is evaluated for significance based on the total number of tests being performed. Specifically, the working $p$-value is calculated as the original $p$-value divided by the number of tests being performed. Implicitly, it assumes that these test statistics are independent. So in our current example an overall desired significance level of .05 would translate into individual tests each using a $p$-value threshold of $.05/8 = .0062$. These thresholds could also be used to create wider confidence intervals for each point estimate as we'll see later in Figure 2 which plots the point estimates from the model above along with both uncorrected and Bonferroni-corrected uncertainty intervals corresponding to a nominal .05 significance level. While the standard intervals reject the null hypothesis of no effect of the intervention for seven sites, the multiple-comparisons-adjusted intervals reject the null hypothesis for only four sites.

The Bonferroni correction directly targets the Type 1 error problem but it does so at the expense of Type 2 error. By changing the $p$-value needed to reject the null (or equivalently widening the uncertainty intervals) the number of claims of rejected null hypotheses will indeed decrease on average. While this reduces the number of false rejections, it also will increase the number of instances that the null is not rejected when in fact it should have been. Thus, the Bonferroni correction can severely reduce our power to detect an important effect.

### 2.4 Other classical corrections

Motivated by the shortcomings of the Bonferroni correction, other researchers have proposed more sophisticated procedures. The goal of these methods typically is to reduce the familywise error rate without unduly sacrificing power. A natural way to achieve this is by taking account of the dependence across tests. A variety of such corrections exist that rely upon bootstrapping methods or permutation tests (see, for example, Westfall and Young, 1993)

A more recent class of approaches to this problem focuses not on reducing the familywise error rate (again, the risk of any false positives) but instead on controlling the expected proportion of false positives, or the "false discovery rate" (Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001). The rationale behind the concern for false discovery rate is that the researcher should be more worried about a situation in which many tests show up as statistically significant and an unknown proportion of these are erroneous than a situation in which all but a few tests show up as insignificant. Controlling for the false discovery rate rather than the familywise error rate leads to a less conservative testing procedure with respect to Type 1 error but is more powerful in terms of detecting effects that are real. These methods make particular sense in fields like genetics where one would expect to see a number of real effects amidst a vast quantity of zero effects such as when examining the effect of a treatment on differential gene expression (Grant, et al., 2005). We see these as less relevant in social science applications.

# 3 A different perspective on multiple comparisons

Classical methods typically start with the assumption that the null hypothesis is true – an unhelpful starting point as we discuss below. Moreover, we argue that they fail to model the parameters corresponding to the tests of interest correctly. Our goal in this paper is not to say that we're proposing the optimal statistical method for all circumstances. Rather we present an entirely different perspective on the issue and its implications and argue that, when viewed from a Bayesian perspective, many of these problems simply disappear.

## 3.1 Abandoning the Type 1 error paradigm

The classical perspective worries primarily about Type 1 errors and we argue that these should not be the focus of our concern. Suppose we've established the following two hypotheses regarding our site-specific treatment effects $\tau_j$ for $j = 1, \ldots, J$: $H_0^j : \tau_j = 0$, and $H_A^j : \tau_j \neq 0$. A primary concern from the classical multiple comparisons perspective is that we might erroneously accept $H_A^j$ when, in fact, $H_0^j$ is true (Type 1 error). But do we ever believe that $\tau_j$ exactly equals zero? What is the practical import of such a test? Similarly, a Type 2 error occurs when we mistakenly accept the $H_0^j$ that $\tau_j = \tau_k$ when in fact the $H_A^j$ that $\tau_j \neq \tau_k$ is true. Again, under what circumstances do we truly believe that are absolutely no differences between groups? There may be no *practical* differences but this is a distinct point which we shall discuss shortly. Moreover, if true effects are zero, we don't want anything close to a 5% chance of finding statistically significant results.

A more serious concern might be that we make a claim that $\tau_j > 0$ when in fact $\tau_j < 0$, in other words, we claim that there is a positive effect when in fact the effect is detrimental.

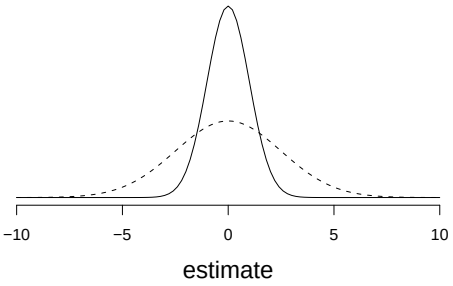


Figure 1: In a particular model, where the null hypothesis happens to be true (or very close to true), the narrow distribution could correspond to estimates of main effects, which have high precision and are thus not claimed to be large, with the wide distribution representing interactions, which have low precision and thus, ironically, will be estimated to be very large in the rare cases when they happen by chance to be statistically significant."

A similar phenomenon occurs if we claim that $\tau_j > \tau_k$ when in fact $\tau_j < \tau_k$; for instance we claim that Miami had a larger treatment effect than New York when in fact the reverse is true. These are both examples of what is referred to as "Type S" error (Gelman and Tuerlinckx, 2000) and are of greater concern.

But in policy analysis, there is also a fair bit of concern with examples where the differences might actually be very close to zero: for example, comparing different educational interventions, none of which might be very effective. Here we would want to be thinking about "Type M" (magnitude) errors: saying that a treatment effect is near zero when it is actually large, or saying that it's large when it's near zero (Gelman and Tuerlinckx, 2000). In that setting, underpowered studies present a real problem. Consider an example in which the true effect was zero (or very close to zero). An ironic property about effect estimates with relatively large standard errors is that they are more likely to produce effect estimates that are larger in magnitude than effect estimates with relatively smaller standard errors, as can be seen in Figure 1. Thus, when we switch from examining main effects to subgroup effects, for example, we automatically increase our probability of seeing large estimates and tricking ourselves into thinking that something is going on. There is a tendency sometimes towards downplaying a large standard error (which might increase the $p$-value of their estimate) by pointing out that, however, the *magnitude* of the estimate is quite large. In fact, this "large effect" is likely a *byproduct* of this standard error. Bayesian modeling will help us here too, as we shall see below.

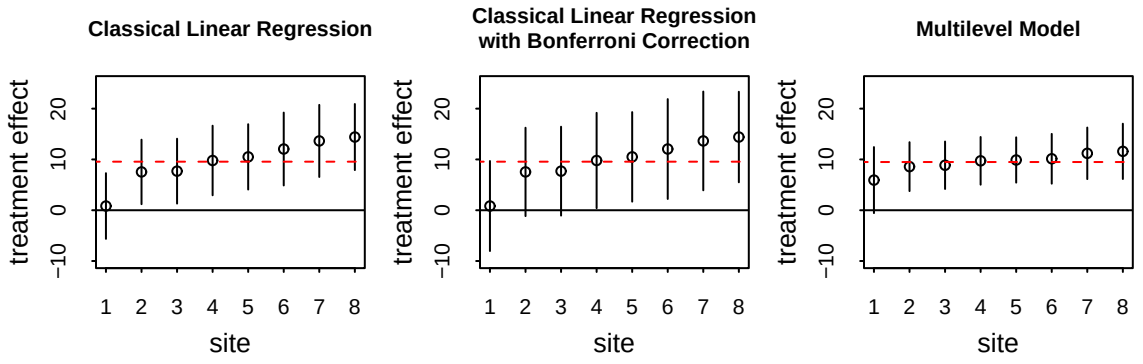


Figure 2: Plots of treatment effect point estimates and 95% intervals across the eight Infant Health and Development Program sites. The left panel display classical estimates from a linear regression. The middle panel displays the same point estimates as in the left panel but the confidence intervals have been adjusted to account for a Bonferroni correction. The right panel displays 95% intervals and means from the posterior distributions for each of the eight site-specific treatment effects generated by fitting a multilevel model.

### 3.2 Multilevel modeling in a Bayesian framework

More strongly, we claim that when viewed within a Bayesian framework, many of these problems disappear, or in the case of Type S and Type M errors, are at least substantially ameliorated. A relatively simple multilevel model would be appropriate in this setting. In this model we might assume that the individuals in a common site experience the same effect on age 3 test scores, such as

$$y_i \sim \mathrm{N}(\gamma_{j[i]} + \delta_{j[i]} P_i, \sigma^2),$$

Here $\delta_{j[i]}$ is the parameter for the treatment effect corresponding to person $i$'s site (indexed by $j$). Given that the programs and types of children are by design similar, is also seems reasonable to assume that these effects vary across sites according to a common distribution, such as

$$\delta_j \sim \mathrm{N}(\mu, \omega^2).$$

We've also allowed the intercept, $\gamma$, to vary across sites in a similar manner. It does not seem a strong assumption to think of these as realizations from a common distribution and this addition should strengthen our model. Additionally, our Bayesian analysis requires us to specify prior distributions for the parameters $\mu$, $\sigma$, and $\omega$. However (particularly for this kind of simple model) it is not difficult to choose priors to be so uninformative that they have little to no impact on our inferences. Finally, we could (and should, in a real analysis) easily add other predictors to the model to increase our predictive power but have refrained from doing so to focus on the primary issues.

**Partial pooling.** This model can be thought of as a compromise between two extremes. One extreme, complete pooling, would assume the treatment effects are the same across all sites, that is, $\delta_j = \delta$, for all $j$. The other extreme, no pooling, would estimate treatment effects separately for each site. The compromise found in the multilevel model is often referred to as *partial pooling*. Figure 2, graphically illustrates this compromise by plotting the multilevel intervals next to the classical estimates and intervals (with and without Bonferroni corrections) for comparison. The horizontal dashed line displays the complete pooling estimate. We also display a horizontal solid line at 0 to easily discern which estimates would be considered to be statistically significant. This process leads to point estimates that are closer to each other (and to the "main effect" across all sites) than the classical analysis. Rather than inflating our uncertainty estimates, which doesn't really reflect the information we have regarding the effect of the program, we shift the point estimates in ways that do reflect the information we have. It has been recognized since the pioneering work of James and Stein (1960) and Efron and Morris (1975) that partial pooling can lead to estimates with better properties (for instance lower mean squared error) than traditional estimators.

**The intuition.** Why does partial pooling make sense at an intuitive level? Let's start from the basics. The only reason we have to worry about multiple comparisons issues is because we have uncertainty about our estimates. If we knew the "true" treatment effect in each site we wouldn't be making any probabilistic statements to begin with – we would just know the true sign and true magnitude of each (and certainly then whether or not each was really different from 0 or from each other). Classical inference in essence uses only the information in each site to get the treatment effect estimate in that site and the corresponding standard error. A multilevel model however recognizes that this site-specific estimate is actually ignoring some important information – the information provided by the other sites. While still allowing for heterogeneity across sites, the multilevel model also recognizes that since all the sites are measuring the same phenomenon it doesn't make sense to complete ignore what has been found in the other sites. Therefore each site-specific estimate gets "shrunk" or pulled towards the overall estimate. The greater the uncertainty in a site, the more it will get pulled towards the overall estimate. The less the uncertainty in a site, the more we trust that individual estimate and the less it gets shrunk.

**Model fitting.** One barrier to more widespread use of multilevel models is that researchers aren't always sure how to fit such models. We often recommend fitting multilevel models in a fully Bayesian way using a software package such as Bugs (as described in

detail in Gelman and Hill, 2007). However many simple models can be fit quite well using packages that have been built in (or can be easily installed into) existing software packages. For instance the model above can be fit easily in R

```
ihdp.fit <- lmer(y~ treat+(1+treat|state))
```

and further functions exist to help the user sample from the posterior distribution for each site-specific treatment effect (or any other parameter from the model or functions thereof; see Gelman and Hill, 2007). Similar options for fitting the model are available in Stata and SAS as well (see Appendix C of Gelman and Hill, 2007).

**Other motivations for multilevel models.** The multiple comparisons problems arise frequently in research studies in which participants have been clustered because of interest in examining differences across these program sites, schools, cities, etc. However, arguably, data from these types of studies should already be fit using a multilevel model.

# 4 Examples illustrating the relevance and lack of relevance of multiple comparisons in different settings

## 4.1 Comparing the impact of multiple treatments: effects of electromagnetic fields at 38 frequencies

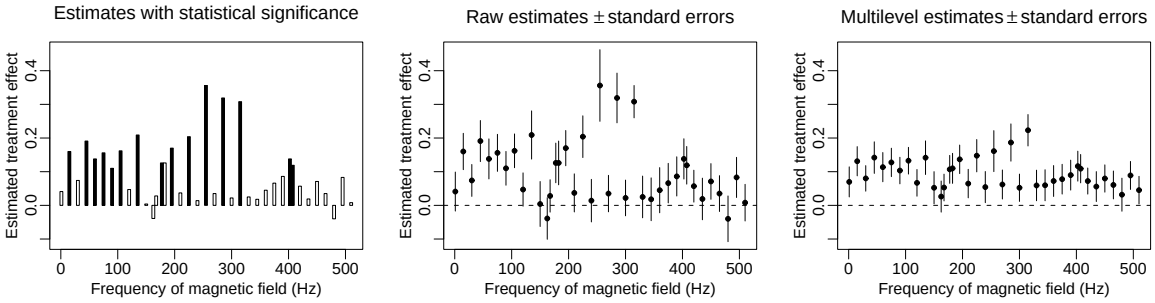


Figure 3: (a) Estimated effects of electromagnetic fields on calcium efflux from chick brains, shaded to indicate different levels of statistical significance, adapted from Blackman et al. (1988). A separate experiment was performed at each frequency. (b) The same results presented as estimates ± standard errors. The first plot, with its emphasis on statistical significance, is misleading. (c) Multilevel estimates and standard errors, partially pooled from the separate estimates displayed in the center graph. The standard errors of the original estimates were large, and so the multilevel estimates are pooled strongly toward the common mean estimate of approximately 0.1.

In the wake of concerns about the health effects of low-frequency electric and magnetic fields, an experiment was performed to measure the effect of electromagnetic fields

at various frequencies on the functioning of chick brains (Gelman and Hill, 2007). At each of several frequencies of electromagnetic fields (1 Hz, 15 Hz, 30 Hz, ..., 510 Hz), a randomized experiment was performed to estimate the effect of exposure on calcium efflux, compared to a control condition of no electromagnetic field. The researchers reported, for each frequency, the estimated treatment effect (the average difference between treatment and control measurements) and the standard error[2]. Results in the original article were summarized in terms of $p$-values (see Figure 3a) but are more informatively displayed as confidence intervals, as in Figure 3b.

The confidence intervals at different frequencies in Figure 3b overlap substantially, which implies that the estimates could be usefully pooled using a multilevel model. The raw data from these experiments were not available,[3] so we analyzed the estimates, which we shall label as $y_j$ for each subexperiment $j$. Because the chicken brains were randomly assigned to the treatment and control groups, we can assume the $y_j$'s are unbiased estimates of the treatment effects $\theta_j$; and because the sample sizes were not tiny, it is reasonable to assume that the estimation errors are approximately normally distributed; thus, $y_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$. Our default model for the treatment effects is simply $\theta_j \sim \mathrm{N}(\mu_\theta, \sigma_\theta^2)$. If we assume each $\sigma_j$ is known and equal to the standard error of the estimate $y_j$, we can easily perform inference about the $\theta_j$'s (as well as the hyperparameters $\mu_\theta$ and $\sigma_\theta$)

Figure 3c displays the inferences for the treatment effects $\theta_j$, as estimated from the multilevel model. The inferences shown here represent partial pooling of the separate estimates $y_j$ toward the grand mean $\mu_\theta$. More generally, the multilevel model can be seen as a way to estimate the effects at each frequency $j$, without setting "nonsignificant" results to zero. Some of the apparently dramatic features of the original data as plotted in Figure 3a—for example, the negative estimate at 480 Hz and the pair of statistically significant estimates at 405 Hz—do not stand out so much in the multilevel estimates, indicating that these features could be easily explained by sampling variability and do not necessarily represent real features of the underlying parameters.

**Potential expansions of the multilevel model.** This multilevel model can be criticized because it pools all the estimates toward their common mean of 0.1 with an assumed normal distribution for the true $\theta_j$'s. This would not be appropriate if, for example, the treatment had a positive effect at some frequencies and a zero effect at others. Or, as hypothesized by the authors of the original study, that the true treatment effects could fall into three

[2]By standard error here we mean $\sqrt{\sigma_T^2/n_T + \sigma_C^2/n_C}$ where $\sigma_T^2$ and $\sigma_C^2$ represent the standard deviations of calcium efflux in the control and treatment groups, respectively, and $n_T$ and $n_C$ denote the corresponding sample sizes.

[3]When we asked the experimenter to share the data with us, he refused. This was disturbing considering this was a government-funded study of public health interest.

groups: a set of large effects near 0.3, a set of moderate effects near 0.15, and a set of zero effects.

We could explore this possibility by fitting a mixture model for the $\theta_j$'s. But the resulting inferences would not differ much from our multilevel analysis that used a normal distribution. The reason that changing the model would not do much is that the uncertainty bounds for the individual estimates are so high. Even if, for example, we fit a model with three clusters of effects, it would not be so clear which points correspond to which cluster. The estimates at 255, 285, and 315 Hz appear to be one cluster, but in fact the point at 255 Hz has a high standard error (see Figure 3b) and could very well belong to a lower cluster, whereas the estimates at 45, 135, 225, and other points are consistent with being in the higher group. Similarly, several of the estimates are not statistically significant (see Figure 3a) but they are almost all positive, and in aggregate they do not appear to be zero. There is certainly no sharp dividing line between the "low" and the "moderate" estimates.

To put it another way: we do not "believe" the estimates in Figure 3a, but neither do we "believe" the separate estimates in Figure 3b, which we know from statistical theory will almost certainly be more variable than the true effects. We would recommend using a multilevel model and estimates as a starting point for further analysis of these data.

**Relevance to multiple comparisons.** One of the risks in statistical analysis of complex data is overinterpretation of patterns that could be explained by random variation. Multilevel modeling can reduce overinterpretation. For example, Figure 3a shows a dramatic pattern of three points that stand apart from all the others, but the multilevel estimates in Figure 3c show these to be part of a larger group of relatively large effects for frequencies up to 300 Hz or so. The partial pooling has revealed the fragility of the patterns in the raw data (or in the no-pooling estimates), which can be explained by sampling variability.

### 4.2 Comparing average test scores across all U.S. states

This next example illustrates how these issues play out in a situation in which all pairwise comparisons across groups are potentially of interest. Figure 4 shows a graph from a National Center for Education Statistics (1997) report, ordering all states based on average scores on the National Assessment of Educational Progress (NAEP) fourth-grade mathematics test; Bonferroni corrections have been performed and statistically significant comparisons have been shaded. In theory, this plot allows us to answer questions such as, Does North Carolina have higher average test scores than neighboring South Carolina? This information could be displayed better (Wainer, Hambleton, and Meara, 1999, Almond et al., 2000), and maybe should not be displayed at all (Wainer, 1986,) but here our concern

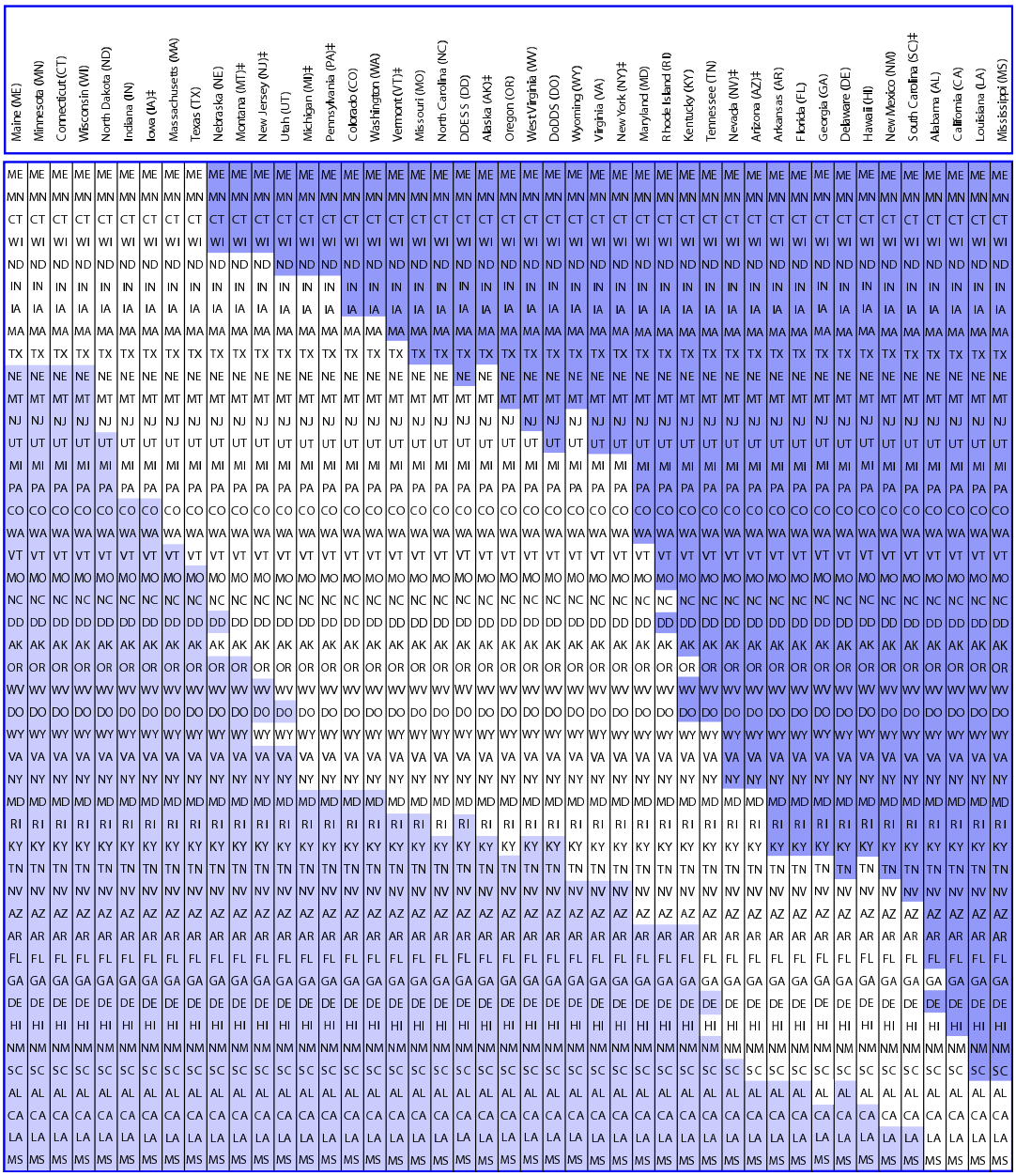


Figure 4: Graph from the National Center for Education Statistics (1997) report comparing average math test scores of students in different states. Shaded comparisons represent differences that are statistically significant after a multiple comparisons correction. The full table printed in the report shows all paired comparisons. We argue that these multiple-comparisons adjustments make no sense, as they are based on the irrelevant model in which true population differences are exactly zero.

is with the formulation as a multiple comparisons problem.

**Concerns with the classical multiple comparisons display.** Here is a situation in which most classical multiple comparisons adjustments, such as the Bonferroni adjustment that was used, will not be appropriate because we know ahead of time that the null hypothesis (zero average differences between states) is false, so there is no particular reason to worry about the Type 1 error rate. Therefore, any motivation for multiple comparisons then rests either on (a) wanting more than 95% of the 95% intervals to contain the true values, or (b) wanting a lower Type S error rate, in other words, we want to minimize the chance of, for instance, stating that New Jersey has higher average test scores than Pennsylvania when, in fact, the opposite is the case.

With regard to 95% intervals, we will do better using multilevel modeling, either on the raw state averages from any given year or, even better, expanding the model to include state-level predictors and test scores from other years. If Type S error rates are a concern, then, again, a multilevel model will more directly summarize the information available.

Of course the objection may be raised that although we may accept that the true differences are not *exactly* zero, but what about the null hypothesis that they are *nearly zero*? Our reply is that this weaker version of classical hypothesis testing doesn't work here either. One way to see this is that the data used to create Figure 3 clearly reject either of these null hypotheses. But the multiple comparisons procedures just plug along ignoring this information, widening the confidence intervals as new comparisons are added.

**Multilevel model and corresponding display of comparisons.** As an alternative, we fit a multilevel model: $y_j \sim \mathrm{N}(\alpha_j, \sigma_j^2)$, where $j = 1, \ldots, J$ are the different states, and $y_j$ is the average fourth grade mathematics score for the students who took the test in state $j$.[4] The parameters $\alpha_j$ represent the true means in each state—that is, the population averages that would be obtained if all the students in the state were to take the test. We model these population averages with a normal distribution: $\alpha_j \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$. Finally, we assign uniform prior distributions to the hyperparameters $\mu_\alpha, \sigma_\alpha, \sigma_j$.

One advantage of working within the Bayesian multilevel modeling paradigm in which models are fit using simulation is that the output is easy to manipulate in order to examine whatever functions of the parameters are of primary interest. In this case, based on the fitted multilevel model, we simulate $S = 1000$ draws of state effect parameters to construct a

[4] We recognize that this model could be improved, most naturally by embedding data from multiple years in a time series structure. The ability to include additional information in a reliable way is indeed a key advantage of multilevel models; however, here we chose a simple model because it uses no more information than was used in the published tables.

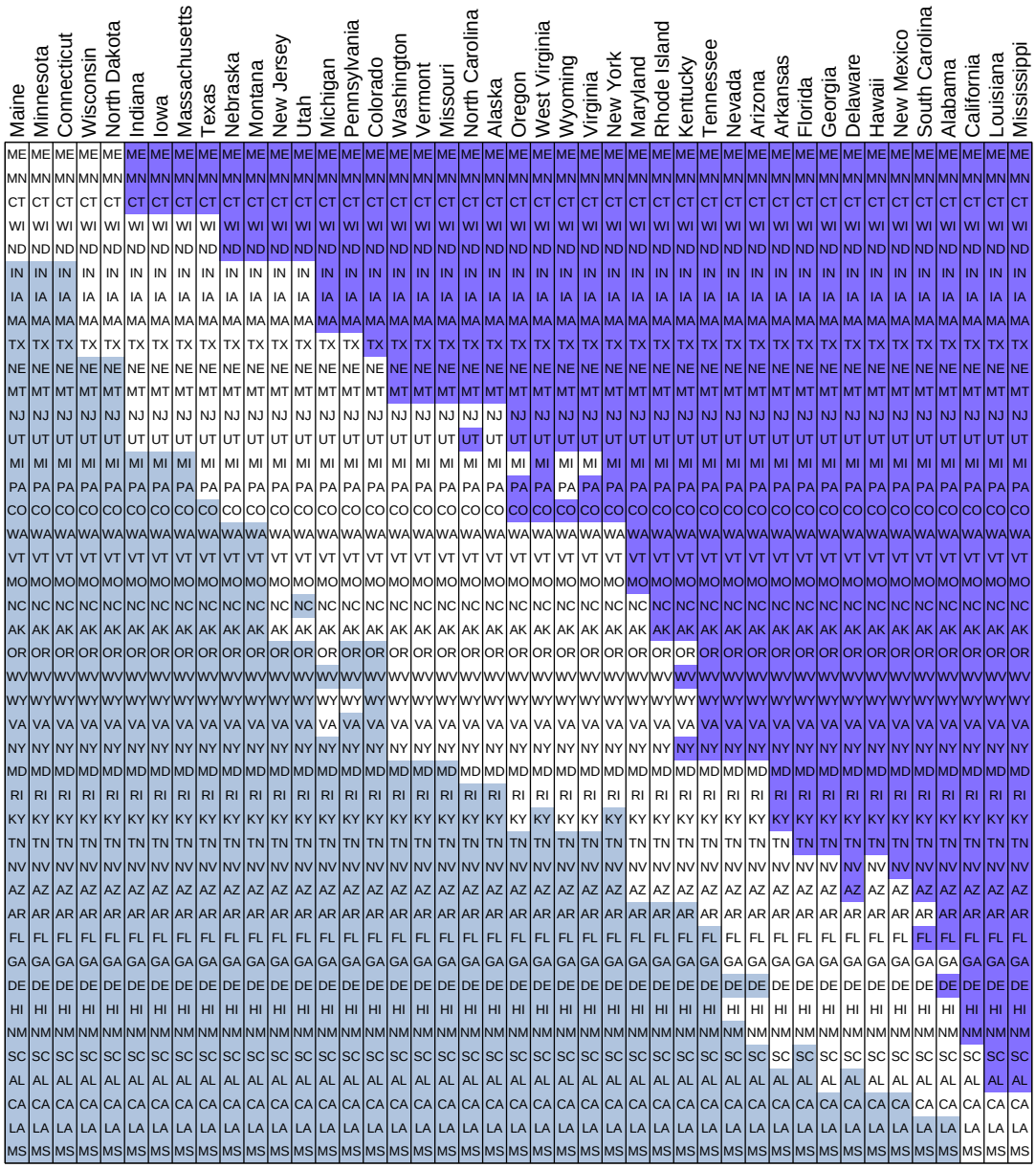


Figure 5: Posterior inference of National Center for Education Statistics (1997) report from a fitted multilevel model. Compared to the classical multiple comparisons summaries in Figure 4, these Bayesian intervals are more confident in making compairsons. This is because the states are far apart in this dataset, thus little pooling is done and the classical multiple comparisons adjustments are overkill.

posterior interval for the difference in true means for each pair of states. For the purpose of comparing to the classical approach, we set a 0.05 cutoff: for each pair of states $j, k$, we check whether 95% or more of the simulations show $\alpha_j > \alpha_k$. If so—that is, if $\alpha_j > \alpha_k$ for at least 950 of the 1000 simulations—we can claim with 95% confidence that state $j$ outperforms state $k$. We plot the results in Figure 5. States that have effects that are significantly lower are shaded with light blue, the ones which are higher are shaded with darker blue, and ones that are not significantly different are left as white. Simulation variability adds some noise to this plot, as can be seen, for example, in some of the non-monotonic patterns of the comparisons across states.

Compared to the classical multiple comparisons summaries in Figure 4, the multilevel estimates in Figure 5 are more informative, with more claims with confidence and fewer cells in the central region where comparisons are not statistically significant. The classical procedure overcorrects for multiple comparisons in this setting where the true differences between states are large. (At the other extreme, classical procedures undercorrect for multiple comparisons when true differences are small, as we discuss in Section 4.3.

The classical multiple comparisons procedure as shown in Figure 4 does not "listen" to the data. When there is evidence for a multiple comparisons problem, our procedure makes corrections. When there is no evidence for a multiple comparisons problem our procedure is similar to the direct inference without a multiple comparisons correction.

### 4.3 SAT coaching in 8 schools

Rubin (1981) discusses an example (reprised in chapter 5 of Gelman et al., 2003) of a meta-analysis of randomized experiments of coaching for the Scholastic Aptitude Test (SAT) in eight high schools in New Jersey. This example is notable as one of the first fully Bayesian analysis of a hierarchical model and also because there was no evidence in the data of differences between the treatment effects in the different schools. (And, in fact, the total estimated effects are small.) The first two columns of numbers in Figure 6 give the data. Just to get a sense of the variation, the standard deviation of the eight school estimates is 10, which is of the same order as the standard errors.

**Classical and Bayesian analysis.** This is the sort of situation where one might worry about multiple comparisons. (In the actual data in Figure 6, none of the comparisons is statistically significant, but as we discuss below, they could be in a replication of the study.)

The hierarchical Bayesian analysis of Rubin (1981) has no multiple comparisons problems, however. The group-level variance is estimated to be low—the marginal maximum likelihood or posterior mode estimate is zero, and the Bayesian analysis averages over the

| School | Raw estimate of treatment effect, $y_j$ | Standard error of raw effect estimate, $\sigma_j$ | Bayes posterior mean | Bayes posterior sd |
|---|---|---|---|---|
| A | 28 | 15 | 11 | 8 |
| B | 8 | 10 | 7 | 6 |
| C | −3 | 16 | 6 | 8 |
| D | 7 | 11 | 7 | 7 |
| E | −1 | 9 | 5 | 6 |
| F | 1 | 11 | 6 | 7 |
| G | 18 | 10 | 10 | 7 |
| H | 12 | 18 | 8 | 8 |

Figure 6: First two columns of numbers: Data from the 8-schools experiment of Rubin (1981). A separate randomized experiment was conducted in each school, and regression analysis gave separate treatment effect estimates (labeled as $y_j$ above) and standard errors (labeled as $\sigma_j$). Effects are on the scale of points in the SAT-Verbal test (which was scored from 200 to 800). An effect of 8 points corresponds to approximately 1 additional test item correct.
Last two columns: Posterior mean and standard deviation of treatment effects, as estimated from a Bayesian multilevel model. The evidence is that the effects vary little between schools, hence the estimates are pooled strongly toward the common mean. None of the comparisons from the Bayesian inference are even close to statistically significant.

posterior distribution, which is largely below 10—and as a result the Bayes estimates are pooled strongly toward the common mean.

**Simulation study with small effects.** To get further insight into this example, we perform repeated simulations of a world in which the true treatment effects in different schools come from a normal distribution with standard deviation 5 (a plausible estimate given the data in Figure 6). For each replication, we simulate eight true values $\theta_1, \ldots, \theta_8$ from this distribution, then simulate data $y_1, \ldots, y_8$ from the eight separate normal distributions corresponding to each $\theta_j$. The standard deviations $\sigma_j$ for each of these distributions is given by Figure 6. Note that relative to the within group standard deviations the between group standard deviation of 5 is small. We then performed both classical and hierarchical Bayesian analyses. For each analysis, we computed all $(8\dot{7})/2 = 28$ comparisons and count the number that are statistically significant (that is, where the difference between the estimates for two schools is more than 1.96 times the standard error for the difference), and of these, we count the number that have the correct sign.

We performed 1000 simulations. Out of these simulations, 6.7% of the classical intervals were statistically significant and, of these, 64% got the sign of the comparison correct. Multiple comparisons corrections are clearly necessary here if we want to avoid making

unreliable statements. By comparison, only 0.5% of the Bayesian intervals are statistically significant (with 73% getting the sign of the comparison correct). The shrinkage of the Bayesian analysis has already essentially done a multiple comparisons correction.

To look at it another way: the classical estimates found at least one statistically significant comparison in 60% of our 1000 simulations. In the Bayesian estimates, this occurred only 6% of the time. The Bayesian analysis here is using a uniform prior distribution on the hyperparameters—the mean and standard deviation of the school effects—and so it uses no more information than the classical analysis. As with a classical multiple comparisons procedure the Bayesian inference recognizes the uncertainty in inferences and correspondingly reduces the number of statistically significant comparisons.

**Simulation study with large effects.** To get a sense of what happens when effects are more clearly distinguishable, we repeat the above simulation but assume the true treatment effects come from a distribution with standard deviation 10. This time, 12% of the classical comparisons are statistically significant, with 88% of these having the correct sign. From the Bayesian analysis, 4% of the comparisons are statistically significant, with 96% of these having the correct sign. Whether using classical multiple comparisons or Bayesian hierarchical modeling, the price to pay for more reliable comparisons is to claim confidence in fewer of them.

### 4.4 Teacher and school effects in NYC schools

Rockoff (2004) and Kane, Rockoff, and Staiger (2007) analyzed a huge dataset from the New York City school system in order to assess the importance of factors such as educational background, training, and experience in determining the effectiveness of teachers. One of the findings was that variation in teacher effects on student grades was moderately large, about 0.15 standard deviations on a scale in which standard deviation was calculated using test scores for all students of a given grade level in the system.[5]

This study could have been set up as a multiple comparisons problem, trying to get appropriate $p$-values for comparing thousands of teachers or for distinguishing individual teacher effects from zero. Instead, though, the researchers learned from the scale of unexplained variation in teacher effects—the residual group-level variance—that teacher effects are important and are largely not explained by background variables, except for a small improvement in performance during the first decade of a teacher's career.

[5]Despite the observational nature of the data, it was reasonable to attribute these patterns to causality because: (1) the analysis controlled for pretest scores and other student and class-level variables from before the students were assigned to the particular teachers; (2) estimated teacher effects were persistent from year to year, with top-performing teachers staying in the top year after year; and (3) teachers have essentially no power to select or remove students from their classes.

### 4.5 Fishing for significance: do beautiful parents have more daughters?

In an analysis of data from 2000 participants in the U.S. adolescent health study, Kanazawa (2007) found that more attractive people were more likely to have girls, compared to the general population: 52% of the babies born to people rated "very attractive" were girls, compared to 44% girls born to other participants in the survey. The difference was statistically significant, with a $t$-value of 2.43. However, as discussed by Gelman (2007), this particular difference—most attractive versus all others—is only one of the many plausible comparisons that could be made with these data. Physical attractiveness in the survey used by this paper was measured on a five-point scale. The statistically significant comparison was between category 5 ("very attractive") vs. categories 1–4. Other possibilities include comparing categories 4–5 to categories 1–3 (thus comparing "attractive" people to others), or comparing 3–5 to 1–2, or comparing 2–5 to 1. It is not a surprise that one of this set of comparisons comes up statistically significant.

This study is a good example of a situation in which classical multiple comparisons adjustments are reasonable. Actual sex ratio differences tend to be very small—typically less than 1 percentage point—and so the null hypothesis is approximately true here. A simple Bonferroni correction based on the number of possible comparisons (20) would change the critical value from .05 to .0025 in which case the finding above (with $p$-value of .015) would not be statistically significant.

With a properly designed study, however, multiple comparisons adjustments would not be needed here either. To start with, a simple analysis (for example, linear regression of proportion of girl births on the numerical attractiveness measure) should work fine. The predictor here is itself measured noisily (and is not even clearly defined) so it would probably be asking too much to look for any more finely-grained patterns beyond a (possible) overall trend. More importantly, the sample size is simply too low here, given what we know from the literature on sex-ratio differences. From a classical perspective, an analysis based on 2000 people is woefully underpowered and has a high risk of both Type S and Type M errors.

Alternatively, a Bayesian analysis with a reasonably prior distribution (with heavy tails to give higher probability to the possibility of a larger effect) reveals the lack of information in the data (Gelman and Weakliem, 2007). In this analysis the probability that the effect is positive is only 58%.

### 4.6 Examining impacts across subgroups

We build on our initial example to illustrate how a multi-site analysis could be expanded to accommodate subgroup effects as well. The most important moderator in the Infant Health

and Development Program study was the birth-weight group designation. In fact there was reason to believe that children in the lower low-birth-weight group might respond differently to the intervention than children in the higher low-birth-weight group. We expand our model to additionally allow for differences in treatment effects across birth weight group,

$$y_i \sim \mathrm{N}(\gamma_{j[i]} + \delta^{\mathrm{L}}_{j[i]} P_i(1 - B_i) + \delta^{\mathrm{H}}_{j[i]} P_i B_i, \sigma^2),$$

Here the treatment effect corresponding to person $i$'s site (indexed by $j$) differs depending on if the child belongs to the lower low-birth-weight group $\delta^{\mathrm{L}}_{j[i]}$ or the higher low-birth-weight group $\delta^{\mathrm{H}}_{j[i]}$. This time each of these sets of parameters gets it's own distribution

$$\begin{aligned} \delta^{\mathrm{L}}_j &\sim \mathrm{N}(\mu_{\mathrm{L}}, \omega^2_{\mathrm{L}}), \text{ and} \\ \delta^{\mathrm{H}}_j &\sim \mathrm{N}(\mu_{\mathrm{H}}, \omega^2_{\mathrm{H}}). \end{aligned}$$

In this case we allow the treatment effects for the lower and higher low-birth-weight children to be correlated with each other with correlation parameter denoted $\rho$. Again we have allowed the intercept, $\gamma$, to vary across sites and have specified prior distributions for the hyperparameters that should have little to no impact on our inferences.

Figure 7 plots some of the results from this model. The results for the lower low-birth-weight group are quite volatile in the classical setting, where the information we have about the relationship between the sites is ignored. The Bonferroni correction serves only to reinforce our uncertainty about these estimates. On the other hand, the results from the Bayesian multilevel model for this group have been partially pooled towards the main effect across groups and thus are less subject to the idiosyncracies that can arise in small samples (the sample sizes across these sites for this group range from 18 to 43). The results for the higher low-birth-weight children are more stable for all analyses relative to those for the lower low-birth-weight group due in part at least to the larger sample sizes (they range from 64 to 100). The results from the Bayesian multilevel analysis though are the most stable and lead to substantively different conclusions in that not one of the 95% intervals covers zero. In comparison, in the classical analyses, one site fails to reject the null hypothesis when we don't adjust for multiple comparisons and when we do several more are on the borderline of significance.

## 5 Multiple outcomes and other challenges

Similar issues arise when researchers attempt to evaluate the impact of a program on many different outcomes. If you look at enough outcomes eventually one of them will appear to demonstrate a positive and significant impact, just by chance. In theory, multilevel models

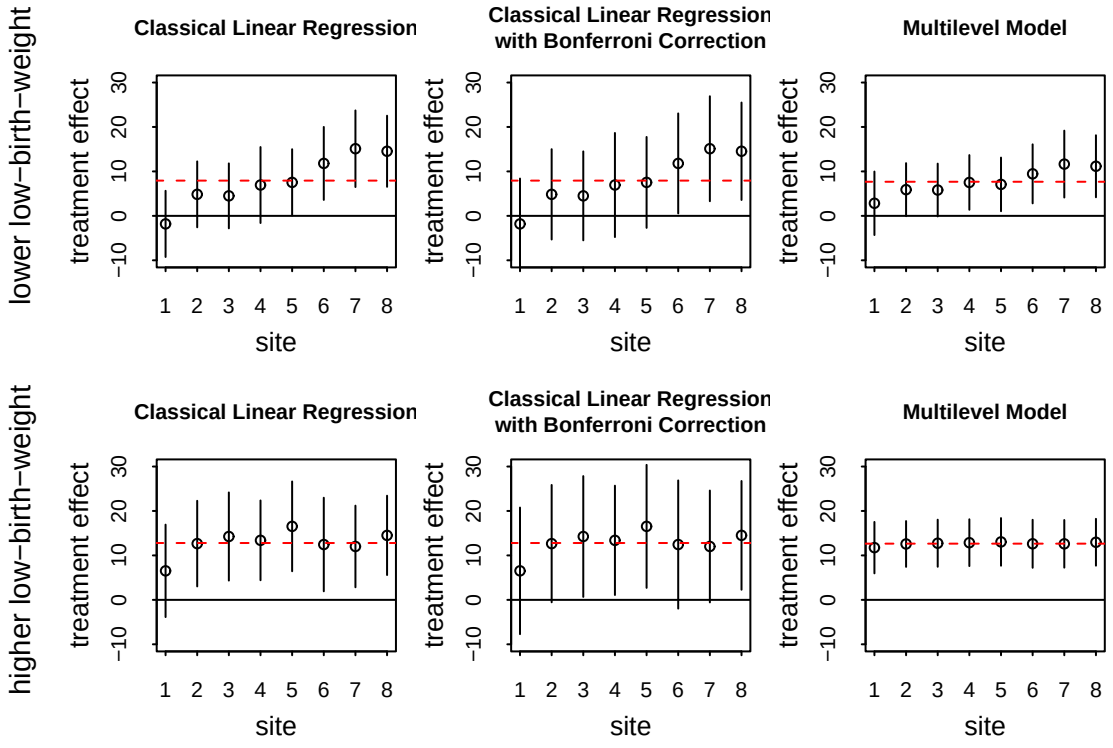


Figure 7: Plots of treatment effect point estimates and 95% intervals across the eight Infant Health and Development Program sites now broken down by birth-weight group as well. The left panel display classical estimates from a linear regression. The middle panel displays the same point estimates as in the left panel but the confidence intervals have been adjusted to account for a Bonferroni correction. The right panel displays 95% intervals and means from the posterior distributions for each of the eight site-specific treatment effects generated by fitting a multilevel model.

can be extended to accommodate multiple outcomes as well. However this often requires a bigger methodological and conceptual jump. The reason that multilevel modes were such a natural fit in the examples described above is because all the estimated groups effects were estimates of the same phenomenon. Thus it was reasonable to assume that they were *exchangeable*. Basically this means that we could think of these estimates as random draws from the same distribution without any a priori knowledge that one should be bigger than another (if we had such information then we would consider that bigger estimate to be a draw from a different distribution). This is not always a reasonable assumption for a group of outcomes.

On the other hand, there are some situations when modeling multiple outcomes simultaneously within a common multilevel model might be fairly natural. For instance, sometimes researchers acquire several different measures of the same phenomenon (such as educational achievement, behavioral problems, or attitudes towards related issues). It is also common to measure the same attribute at several different time points over the course of the study. This might require a slightly more complicated modeling strategy to account for trends over time, but is otherwise a reasonable choice. But if many outcomes have been measured across several disparate domains multilevel modeling might not be the optimal choice.

More generally, if you are comparing a set of exchangeable items, such as 50 states or 8 schools or 4 machine learning algorithms, it's pretty easy to set up a multilevel model. Some choices need to be made—normal distribution or $t$ distribution, mixture model or not, equal or unequal variances, etc.—but in practice it's not hard to come up with a reasonable model based on one's general understanding of the situation. Even if the items differ in recognizable ways, we can still make them exchangeable by including covariates, for example statewide poverty levels or previous test scores in the NAEP example.

Harder problems arise when modeling multiple comparisons that have more structure. For example, 3 varieties of treatments and subgroups classified by 2 sexes and 4 racial groups. We would not want to model this $2 \times 3 \times 4$ structure as 24 exchangeable groups. We think that, even in these more complex situations, multilevel modeling should and will eventually take the place of classical multiple comparisons procedures. After all, as far as we know, every classical multiple comparisons procedure in existence is exchangeable in the sense of treating all the different comparisons symmetrically. As with in many settings, there is a limit to what can be done by building upon the rickety structure of least squares. The more comparisons we have, the more stable our multilevel models will be.

# 6 Conclusion

Multiple comparisons can indeed create problems and it is useful to address the issue. However, statistical methods must be mapped to the setting. Classical Type 1 and Type 2 errors and false discovery rates are based on the idea that the true effect could really be zero (see Johnstone and Silverman, 2004, and Efron, 2006, for connections between these ideas and hierarchical Bayes methods). Effects that are truly zero (not just "small") can make sense in genetics (Efron and Tibshirani, 2002) but not in social science. We prefer to frame the issue in terms of Type S or Type M errors.

Therefore for a social science or program evaluation setting we do not recommend classical fixes that alter $p$-values or (equivalently) make confidence intervals wider. Instead, we prefer multilevel modeling, which shifts point estimates and their corresponding intervals closer to each other (that is, performs partial pooling) where necessary—especially when much of the variation in the data can be explained by noise. By comparison, classical intervals can have Type S error rates as high as 50%; then, multiple comparisons can pose a real threat.

Functions for fitting multilevel models are now available in many statistical software packages; therefore, implementing our suggestions should not be overly burdensome. This should yield better results than the simplest multiple comparisons corrections and should not pose a greater burden than performing one of the fancier types of classical multiple comparisons corrections. Moreover, fitting the multilevel model should result in positive externalities of better point estimates.

We recognize that multilevel modeling can prove to be more of a research challenge for complicated structures as discussed above. More research needs to be done in this area. However we believe it is more worthwhile to invest research time and effort towards expanding the multilevel model framework than to invest in classical multiple comparisons adjustments that start from a perspective on the problem that we find to be unhelpful.

# References


Almond, R. G., Lewis, C., Tukey, J. W., and Yan, D. (2000). Displays for comparing a given state to many others. *American Statistician* **34**, 89–93.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29(4)**, 1165–1188.

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.

Efron, B. (2006). Size, power, and false discovery rates. Technical report, Department of Statistics, Stanford University.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70**, 311–319.

Efron, B., and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

Gelman, A. (2007). Letter to the editor regarding some papers of Dr. Satoshi Kanazawa. *Journal of Theoretical Biology* **245**, 597–599.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

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

Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373–390.

Gelman, A., and Weakliem, D. (2007). Of beauty, sex, and power: statistical challenges in estimating small effects. Technical report, Department of Statistics, Columbia University.

Genovese, C., and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B* **64**, 499–517.

Grant, G. R., Liu, J., and Stoeckert, C. J. (2005). A practical false discovery rate approach to identifying patterns of differential experession in microarray data. *Bioinformatics* **21**, 2684–2690.

Greenland, S., and Robins, J. M. (1991). Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* **2**, 244–251.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* **99**, 609–619.

Harris, W. S., Gowda, M., Kolb, J. W., Strychacz, C. P., Vacek, J. L., Jones, P. G., Forker,

A., O'Keefe, J. H., and McCallister, B. D. (1999). A randomized, controlled trial of the effects of remote, intercessory prayer on outcomes in patients admitted to the coronary care unit. *Archives of Internal Medicine* **159**, 2273–2278.

Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*. London: Chapman and Hall.

Infant Health and Development Program (1990). Enhancing the outcomes of low-birth-weight, premature infants. A multisite, randomized trial. *Journal of the American Medical Association* **263**, 3035–3042.

James, W., and Stein, C. (1960). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, ed. J. Neyman, 361–380. Berkeley: University of California Press.

Johnstone, I., and Silverman, B. (2004). Needles and straw in a haystacks: empirical Bayes approaches to thresholding a possibly sparse sequence. *Annals of Statistics* **32**, 1594–1649.

Kanazawa, S. (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology*.

Kane, T. J., Rockoff, J. E., and Staiger, D. O. (2007). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association* **94**, 1372–1381.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association* **79**, 393–398.

National Center for Education Statistics (1997). 1996 NAEP comparisons of average scores for participating jurisdictions. Washington, D.C.: Government Printing Office.

Rockoff, J. (2004). The impact of individual teachers on student achievement: evidence from panel data. *American Economic Review*, Papers and Proceedings, May.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–401.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., and Lunn, D. (1994, 2002). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England.

Tukey, J. W. (1953). The problem of multiple comparisons. Mimeographed notes, Princeton University.

Wainer, H. (1986). Five pitfalls encountered when trying to compare states on their SAT scores. *Journal of Educational Statistics* **11**, 239–244.

Wainer, H. (1996). Using trilinear plots for NAEP state data. *Journal of Educational Measurement* **33**, 41–55.

Wainer, H., Hambleton, R. K., and Meara, K. (1999). Alternative displays for communicating NAEP results: a redesign and validity study. *Journal of Educational Measurement* **36**, 301–335.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. New York: Wiley.