Journal of Educational Statistics
Winter, 1981, Volume 6, Number 4, pp. 377-400

ESTIMATION IN PARALLEL RANDOMIZED EXPERIMENTS

Donald B. Rubin

Educational Testing Service

Key words: Bayes estimation; coaching experiments; empirical Bayes estimation; simulation of posterior distributions; testing model adequacy

ABSTRACT

Many studies comparing new treatments to standard treatments consist of parallel randomized experiments. In the example considered here, randomized experiments were conducted in eight schools to determine the effectiveness of special coaching programs for the SAT. The purpose here is to illustrate Bayesian and empirical Bayesian techniques that can be used to help summarize the evidence in such data about differences among treatments, thereby obtaining improved estimates of the treatment effect in each experiment, including the one having the largest observed effect. Three main tools are illustrated: 1) graphical techniques for displaying sensitivity within an empirical Bayes framework, 2) simple simulation techniques for generating Bayesian posterior distributions of individual effects and the largest effect, and 3) methods for monitoring the adequacy of the Bayesian model specification by simulating the posterior predictive distribution in hypothetical replications of the same treatments in the same eight schools.

I. INTRODUCTION

Many studies of the effects of treatments consist of several parallel randomized experiments. For example, Alderman and Powers (1979) studied the effects of high school coaching programs on SAT-V (Scholastic Aptitute Test-Verbal

Section) scores in eight schools, each school conducting a separate randomized experiment. As in the Alderman and Powers (1979) study, the estimated effects of the treatments will vary across the parallel experiments. Some of the variation in estimates is due to sampling fluctuation and some variation may be due to real differences among the various experiments (e.g., different kinds of students at the different schools or different efficacies of the coaching programs). The largest observed effect will commonly attract the most attention because it indicates how successful a new program might be; for example, the school with the largest observed effect may be doing something particularly appropriate that could be implemented at other schools. Our discussion in Section II focuses on the largest observed effect and how much we should believe it.

The purpose of this paper is to illustrate techniques that can be used to help summarize the evidence in data about differences among parallel experiments, thereby obtaining improved estimates of the treatment effect in each experiment. These techniques are known by various names, such as empirical Bayes methods, Bayesian hyperparameter models, and James-Stein estimation. The empirical Bayes model and the ideas behind it presented in Section III are not new; the many statistical references to the type of estimation problem considered here include James and Stein (1961), Lindley and Smith (1972), and Efron and Morris (1977). Nor is the application of these ideas entirely new to educational data, M. Novick et al. (1972), Shigemasu (1976), Wang et al. (1977), and Rubin (1980) being recent examples.

However, our analyses of the coaching example highlight important ideas underlying these methods and demonstrate three practical techniques for implementing them. Graphical techniques, presented in Section IV, display sensitivity of estimation to unknown hyperparameters. Simulation techniques, presented in Section V, produce stem-and-leaf posterior distributions of effects. Extensions of the simulation techniques, presented in Section VI, monitor the adequacy of models within a fully Bayesian context. Our results indicate the dangers inherent in empirical Bayes approaches that condition on a particular value of the hyperparameter; similar dangers exist with Bayesian methods that estimate all parameters by a joint posterior mode.

II. THE RESULTS OF EIGHT RANDOMIZED EXPERIMENTS

Alderman and Powers (1979) conducted separate randomized experiments to estimate the effects of coaching on the SAT-V in each of eight high schools. The dependent variable was a special administration of the SAT-V, a standardized multiple choice test administered by the Educational Testing Service and used to help colleges make admissions decisions; the scores can vary between 200 and 800, with mean about 500 and standard deviation about 100. The SAT examinations are designed to be resistant to short term efforts directed specifically toward improving performance on the test; instead they are designed to reflect knowledge acquired and abilities developed over many years of education. Nevertheless, each of the eight schools in this study considered its coaching program to be very successful at increasing SAT scores.

All students in the experiments had already taken the PSAT(Preliminary SAT), and PSAT-M (Math) and PSAT-V were used as covariates. The results of the experiments are summarized in Table I. In each school the estimated coaching effect and its standard error were obtained by an analysis of covariance adjustment appropriate for a completely randomized experiment.

A cursory examination of Table I may at first suggest that some coaching programs have moderate effects (e.g., 18-28 points), most have small effects (e.g., 0-12 points), and two have small negative effects; however, when we take note of the standard errors of these estimated effects, we see that it is difficult statistically to distinguish between any of the experiments; for example, the 95% two-sided confidence intervals all overlap substantially.

The general overlap in the confidence intervals suggests that all experiments may be estimating the same quantity. If they are, we can obtain a better estimate for *each* experiment by pooling the data across all eight experiments to provide one estimate. Under the hypothesis that all experiments have the same effect and produce independent estimates of this common effect, we form the best estimate from the data in Table I by weighting each experiment's estimate inversely proportional to its variance; that is, the best estimate for each experiment is then $\sum_{1}^{8} (\hat{\mu}_i/V_i) \sum_{1}^{8} (1/V_i)$ where $\hat{\mu}_i$ is

TABLE I

Effects of Special Preparation on SAT-V Scores in Eight Randomized Experiments

School	Number of Students Treatment	 Control	Estimated Treatment Effect	Standard Error of Effect Estimate	Residual Variance
A	28	22	28.39	14.9	2415
B	39	40	7.94	10.2	1880
C	22	17	-2.75	16.3	2168
D	48	43	6.82	11.0	2612
E	25	74	-.64	9.4	1623[a]
F	37	35	.63	11.4	2046[a]
G	24	70	18.01	10.4	1841
H	16	19	12.16	17.6	2314

[a] Regression includes a quadratic term for PSAT-V

the estimated treatment effect based on the data in the *ith* experiment and V_i is the square of the standard error for the *ith* experiment, e.g. $\hat{\mu}_1 = 28.39$ and $V_1 = (14.9)^2$. This pooled estimate is 7.9 points and its variance is $[\sum_1^8 (1/V_i)]^{-1} = 17.4$ because the estimates are independent. Thus, assuming all experiments are estimating the same effect, we would estimate this effect to be 7.9 points with standard error equal to 4.2; this leads to the 95% confidence interval (-0.3, 16.0), or approximately 8 ± 8. Supporting this common answer is the fact that the test of the hypothesis that all $\hat{\mu}_i$ are estimating the same quantity yields an F statistic less than one:

$$\frac{1}{7} \sum_1^8 (\hat{\mu}_i - 7.9)^2/V_i \doteq 0.7.$$

Incidentally eight more points on the SAT-V corresponds to about one more test item correct.

One obvious question is whether we should believe the one common answer for all eight experiments or the eight separate answers; or is there some other answer that is better than either alternative? Of particular interest, should we believe that school A's program might be as efficacious as indicated by the estimated effect of +28 points? In other words, if we reran the experiment in school A next year, should we think it equally likely to see an estimated effect above +28 points as below 28 points?

To get a feeling for the natural variation that we would expect across eight studies if the coaching effects in all eight schools were really the same, let us suppose that they are. Specifically, suppose each school's effect is really 8 points with standard error 13 points (the square root of the mean of the eight squared standard errors, i.e. $[\frac{1}{8} \sum_1^8 V_i]^{\frac{1}{2}}$). Then based on the expected values of normal order statistics we would expect the largest of eight effects to be about 26 points, the six next largest to be about 19, 14, 10, 6, 2, and -3 points, and the smallest to be about -9 points. Notice that these expected observed effect sizes are fairly consistent with the actual observed effect sizes (28, 18, 12, 8, 7, 1, -1, -3). Thus, it would appear imprudent to believe that school A really has an effect as large as 28 points.

Of course this rough analysis does not help us much in deciding how large the effect in school A might be, except to suggest that it probably is not as large as 28 points. However, it might not be wise to believe the pooled answer either, because it suggests that a repeated experiment in school A would be equally likely to yield an estimated effect below 7.9 as above 7.9. The rest of this paper presents more precise ways to study the question of estimating the eight treatment effects with particular attention to the effect in school A. The key idea is to incorporate useful information in the data from the other schools.

III. A BAYES/EMPIRICAL BAYES MODEL FOR PARALLEL EXPERIMENTS

Although there appears to be little evidence that the effects in the eight coaching experiments differ, it is of interest to study how sensitive the estimates of coaching effects are to assumptions. For this purpose we turn to a simple Bayes/empirical Bayes model. This section develops the theory, and the next section applies it to the coaching data.

In the *ith* experiment, assume that the estimator $\hat{\mu}_i$ is normal with mean μ_i(the effect being estimated in the *ith* experiment) and variance V_i(the squared standard error):

$$\hat{\mu}_i \sim N(\mu_i, V_i). \qquad (1)$$

This assumption of normality and known standard error is made routinely when we summarize a study by an estimated effect and its standard error, and we will not question its use here. The design (e.g., sample sizes, covariates) and analysis (e.g., checking for outliers) were such that the assumption is justifiable.

Now we add the crucial Bayesian part of the model. Assume that the μ_i are themselves normally distributed with mean μ_* and variance V_*:

$$\mu_i \sim N(\mu_*, V_*). \qquad (2)$$

There are various ways to motivate this assumption. For example, it may be sensible to think of the eight schools as chosen from a large population of schools with approximately normally distributed effects. Theoretical frequentist motivation for assumption (2) comes from the fact that it often leads to estimators with smaller mean squared errors (cf. Efron & Morris, 1977). Theoretical Bayesian motivation focuses on an a priori assumption of exchangeability among the effects of coaching across schools; that is, before seeing any data, there is no reason to include model restrictions that, for example, force the effect in school A to be larger than in school B. Practically, model (2) is a device to obtain answers between the eight independent estimates and the one common estimate, and experience in a variety of contexts (e.g., Efron & Morris, 1977; Rubin, 1980) suggests that such intermediate answers are superior to either extreme. The appropriateness of (2) is discussed further in Section VI.

The important empirical Bayes point in setting up model (2) is that the data in Table I can provide estimates of μ_* and V_*: equations (1) and (2) imply that, given μ_* and V_*, the $\hat{\mu}_i$ are independently normally distributed with common mean μ_* and variances V_i+V_*. Thus, μ_* should be near the center of the eight $\hat{\mu}_i$, and V_* should reflect the extra variability in the $\hat{\mu}_i$ beyond the variability arising from the variances V_i. That is, even if $V_* = 0$ so that all $\mu_i = \mu_*$, as in the discussion of Section II, we would still expect to see variation in the eight observed $\hat{\mu}_i$ because of their sampling variances, V_i. If all experiments were based on infinite samples (i.e., if all $V_i = 0$), then the observed variance of the $\hat{\mu}_i$ would estimate V_*. The point is simply that an estimate of V_* should reflect only the variability in the $\hat{\mu}_i$ beyond that expected from their standard errors. The discussion at the end of Section II suggests that this extra variability may be small.

Suppose for the moment that we knew μ_* and V_*; standard Bayesian calculations under models (1) and (2) (cf. Box and

Tiao, 1973) show that the joint distribution of $(\mu_1,\ldots,\mu_8)$ given $(\hat{\mu}_1,\ldots,\hat{\mu}_8)$, $(V_i,\ldots,V_8)$ and (μ_*,V_*) is the product of eight independent normals, the distribution of μ_i having mean

$$\lambda_i \hat{\mu}_i + (1 - \lambda_i) \mu_* \tag{3}$$

and variance

$$\lambda_i V_i, \tag{4}$$

where

$$\lambda_i = V_*/(V_i + V_*). \tag{5}$$

Equation (3) is widely known in psychometric literature as Kelley's regression estimate (cf. Lord & Novick, 1968, p. 65), and it provides the correct point estimate of μ_i under (1) and (2) because, conditional on μ_* and V_*, the posterior distribution of μ_i is symmetric about expression (3).

Note from equation (5) that if $V_i/V_* = 0$ then $\lambda_i = 1$, whereas if $V_i/V_* = \infty$ then $\lambda_i = 0$. Consequently, if V_* is large relative to V_i, λ_i will be nearly one and then the appropriate estimate of the effect in the *ith* school will be nearly its separate estimate, $\hat{\mu}_i$. If V_* is small relative to V_i, then λ_i will be nearly zero, and the appropriate estimate in the *ith* experiment will be μ_*. For values of V_* between 0 and ∞, the estimates of μ_i will be between $\hat{\mu}_i$ and μ_*.

If we accept models (1) and (2) and the consequential answer given by expressions (3)-(5), the objective becomes to eliminate the dependence on the unknown parameters μ_* and V_*. Suppose that we knew V_*; then, with a diffuse prior distribution on μ_*, the conditional distribution of μ_* given $(\hat{\mu}_1,\ldots,\hat{\mu}_8)$, $(V_1,\ldots,V_8)$, and V_* is normal with mean

$$\Sigma\hat{\mu}_i\lambda_i/\Sigma\lambda_i \quad (=(\Sigma\hat{\mu}_i/V_i)/\Sigma(1/V_i) \quad \text{if } V \ = 0), \tag{6}$$

with variance

$$V_*/\Sigma\lambda_i$$

$$(= (\Sigma V_i^{-1})^{-1} \text{ if } V \ = 0). \qquad (7)$$

With data like ours, estimates are not sensitive to the choice of a diffuse prior distribution for μ_*.

We now can combine the estimates for $(\mu_1,\ldots,\mu_8)$ in expressions (3)-(5), which are conditional on both μ_* and V_*, with the estimate for μ_* in expressions (6)-(7), which is conditional on V_*, to obtain estimates for $(\mu_1,\ldots,\mu_8)$ which are only conditional on V_*. Specifically, the conditional distribution of $(\mu_1,\ldots,\mu_8)$ given $(\hat{\mu}_1,\ldots,\hat{\mu}_8)$, $(V_1,\ldots,V_8)$, and V_* is eight-variate normal with the means of μ_i given by

$$\lambda_i\hat{\mu}_i + (1 - \lambda_i)\ \Sigma\hat{\mu}_i\lambda_i/\Sigma\lambda_i$$

$$(= \Sigma(\hat{\mu}_i/V_i)/\Sigma(1/V_i) \text{ if } V_* = 0), \qquad (8)$$

the variance of μ_i given by

$$\lambda_i V_i + (1 - \lambda_i)^2\ V_*/\Sigma\lambda$$

$$(= (\Sigma V_i^{-1})^{-1} \text{ if } V_* = 0)\ , \qquad (9)$$

(i.e., the variance is the sum of the expectation of the conditional variance, (4), plus the variance of the conditional mean, (3)), and the covariance between μ_i and μ_j is given by

$$(1 - \lambda_i)(1 - \lambda_j)V_*/\Sigma\lambda_i$$

$$(= (\Sigma V_i^{-1})^{-1} \text{ if } V_* = 0). \qquad (10)$$

The remaining issue is how to handle the dependence of (8)-(10) on the unknown parameter V_*. Because answers can be sensitive to the value of V_*, we choose to examine the estimates of μ_i given by (8)-(10) for a variety of values of V_*, with an indication for each value of V_* of how likely it is. Specifically, for each value of V_*, we will calculate its relative likelihood given the observed values of $\hat{\mu}_1,\ldots,\hat{\mu}_8$ and $V_1,\ldots,V_8$. We will derive this likelihood from a Bayesian argument.

Using standard notation for writing conditional distributions (where the conditioning on $V_1,\ldots,V_K$ is implicit and $\hat{\mu} = (\hat{\mu}_1,\ldots,\hat{\mu}_K)$), we have:

$$f(\mu_*|\hat{\mu},V_*)\ f(V_*|\hat{\mu}) \propto f(V_*)\ f(\mu_*|V_*)\ f(\hat{\mu}|\mu_*,V_*),$$

where $f(V_*)f(\mu_*|V_*)$ is the prior distribution of μ_* and V_*. Because $f(\mu_*|V_*)$ is assumed to be proportional to a constant,

$$f(V_*|\hat{\mu}) \propto f(V_*)\ f(\hat{\mu}|\mu_*,V_*)/f(\mu_*|\hat{\mu},V_*).$$

Under models (1) and (2), the $\hat{\mu}_i$ are conditionally independent $N(\mu_*,V_i + V_*)$ given (μ_*,V_*). Also, μ_* given $\hat{\mu}$ and V_* is normal with mean (6) and variance (7). Thus, we have

$$f(V_*|\hat{\mu}) \propto f(V_*)\ \frac{\Pi\ (V_i + V_*)^{-\frac{1}{2}} \exp[-\tfrac{1}{2}\ \Sigma\ (\mu_i - \mu_*)^2/(V_i + V_*)]}{(V_*/\Sigma\lambda_i)^{-\frac{1}{2}} \exp\left[-\ \tfrac{1}{2}\ \dfrac{(\mu_* - \Sigma\hat{\mu}_i\lambda_i/\Sigma\lambda_i)^2}{V_*/\Sigma\lambda_i}\right]}$$

or

$$\frac{f(V_*|\hat{\mu})}{f(V_*)} \propto$$

$$\left(\frac{\Pi\ \lambda_i}{V_*^{K-1}\ \Sigma\ \lambda_i}\right) \exp\left\{-\ \frac{1}{2V_*}\left[\Sigma\lambda_i\hat{\mu}_i^2 - \frac{(\Sigma\hat{\mu}_i\lambda_i)^2}{\Sigma\lambda_i}\right]\right\} \quad \text{if } V_* > 0 \tag{11}$$

$$\left(= \left(\frac{\Pi V_i^{-1}}{\Sigma V_i^{-1}}\right)^{\frac{1}{2}} \exp[\,-\tfrac{1}{2}\ \Sigma(\hat{\mu}_i/V_i)^2 + \tfrac{1}{2}\ (\Sigma\hat{\mu}_i/V_i)^2/\Sigma V_i^{-1}] \quad \text{if } V_* = 0\right).$$

Expression (11) provides the relative likelihood of each value of V_*, that is, the ratio of the posterior probability to the prior probability. This expression can also be derived by integrating out μ_* from the likelihood of (μ_*, V_*) under models (1) and (2). Values of V_* having relatively high likelihood are more plausible given the data than values of V_* with low likelihood. If $f(V_*) \propto$ constant, then expression (11) is also proportional to the posterior distribution of V_*.

The next section shows how to examine expression (11) in conjunction with expressions (8)-(10) to draw inferences about the effects of coaching in the eight experiments. Dempster, Rubin, and Tsutakawa (1981) discuss more general models than the one presented in this section but focus on computational methods for obtaining inferences conditional on the maximum likelihood estimates of V_*, that is, conditional on the value of V_* which maximizes (11).

IV. APPLICATION OF EMPIRICAL BAYES MODEL TO COACHING EXPERIMENT

The dotted line in Figure 1 is the relative likelihood, given by (11), as a function of $S_* = \sqrt{V_*}$ for the coaching data. Equivalently, Figure 1 is the posterior distribution of S_* under models (1) and (2) assuming the prior distribution of (μ_*, S_*) is constant. In this section we focus on inferences that follow solely from the likelihood function interpretation; Section V presents Bayesian inferences that follow from interpreting the dotted line in Figure 1 as the posterior distribution of S_*.

Values of S_* near zero are most plausible; zero is the most likely value, values of S_* larger than 10 are less than half as likely as $S_* = 0$, and values of S_* larger than 25 are less than one-twentieth as likely as $S_* = 0$. Hence, the range for plausible values of S_* is between about 0 and 25, with smaller values more likely than larger values.

The solid lines in Figure 1 give the estimated effects for the eight experiments as a function of S_*, i.e., the estimates provided by expression (8). For most of the likely values of S_*, the estimated effects are relatively

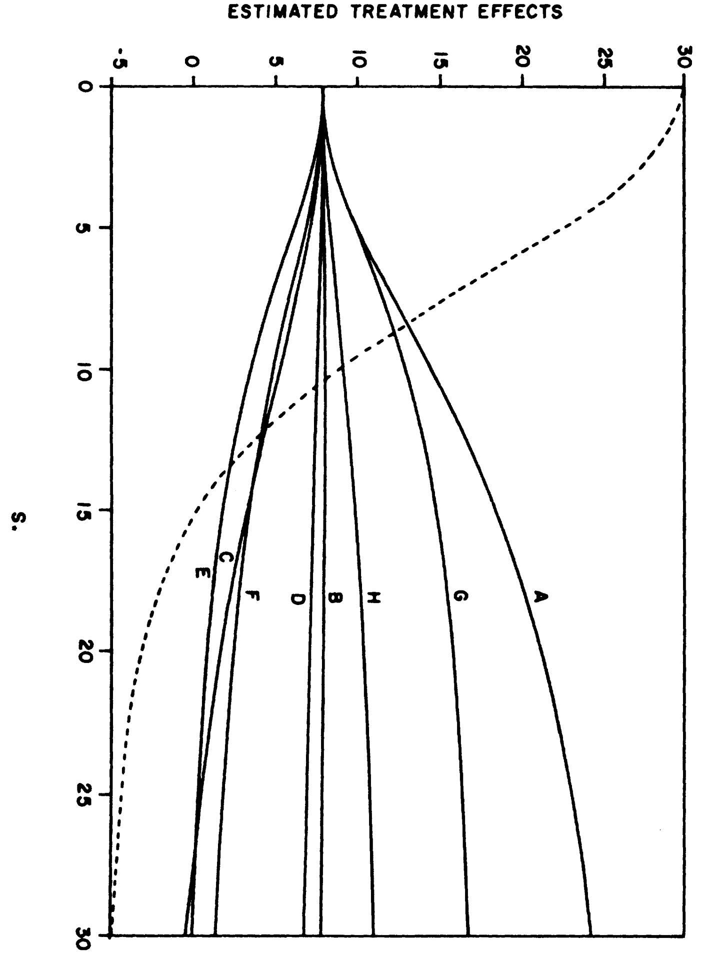

Figure 1: Estimates of Treatment Effects as Functions of S, with Likelihood of S, Superimposed

concentrated about 8 points. For example, at S_* = 10, the estimated treatment effects are about (14.5, 8.1, 5.3, 7.6, 3.6, 5.0, 12.9, 9.2). As S_* becomes larger, the estimated treatment effects become more like the values in Table I. The rising lines in Figure 2 give the standard deviations of the eight estimated effects as a function of S_*, i.e., the square roots of the posterior variances provided by expression (9). For the most likely values of S_*, these standard errors are quite similar to each other; at S_* = 10, the standard errors are (9.1, 7.7, 9.4, 8.0, 7.3, 8.1, 7.8, 9.6). As S_* becomes larger, the standard errors become more like the values in Table I. The falling lines in Figure 2 give the minimum and maximum correlation among the estimates where the correlations are calculated from expressions (9) and (10); when S_* = 0 the estimates are all correlated 1.0, but for larger values of S_* the estimates are nearly uncorrelated. These correlations are needed to calculate the standard deviations of the estimated differences $\mu_i - \mu_j$.

The general conclusion from an examination of Figures 1 and 2 is that an effect in any school as large as 28.4 points is unlikely. For the likely values of S_*, the estimates in all schools are substantially less than 28 points. For example, even at S_* = 10, the probability that the effect in School A is less than 28 points is $\Phi[(28 - 14.5)/9.1] = 93\%$; the corresponding probabilities for the effects being less than 28 points in the other schools are: 99.5%, 99.2% 98.5%, 99.96%, 99.8%, 97% and 98%.

Notice that we do not obtain an accurate summary of the data if we condition on the maximum likelihood estimate of S_*, i.e., S_* = 0. The technique of conditioning on the maximum likelihood estimate of a hyperparameter such as S_* is not an uncommon device. At S_* = 0, however, the inferences are that all experiments have the same size effect, 7.9 points, and the same standard error, 4.2 points. Figures 1 and 2 certainly suggest that this answer represents too much pulling together of the estimates in the eight schools. Bayesian joint posterior modal estimates of $(\mu_*, S_*, \mu_1,\ldots,\mu_8)$ can suffer similar problems.

If the likelihood function of S_* were nearly symmetric about a positive maximum likelihood estimate, then inferences conditional on the maximizing value of S_* and joint posterior modal estimates would be more acceptable; for example, the resulting estimates of the μ_i might result in reasonable

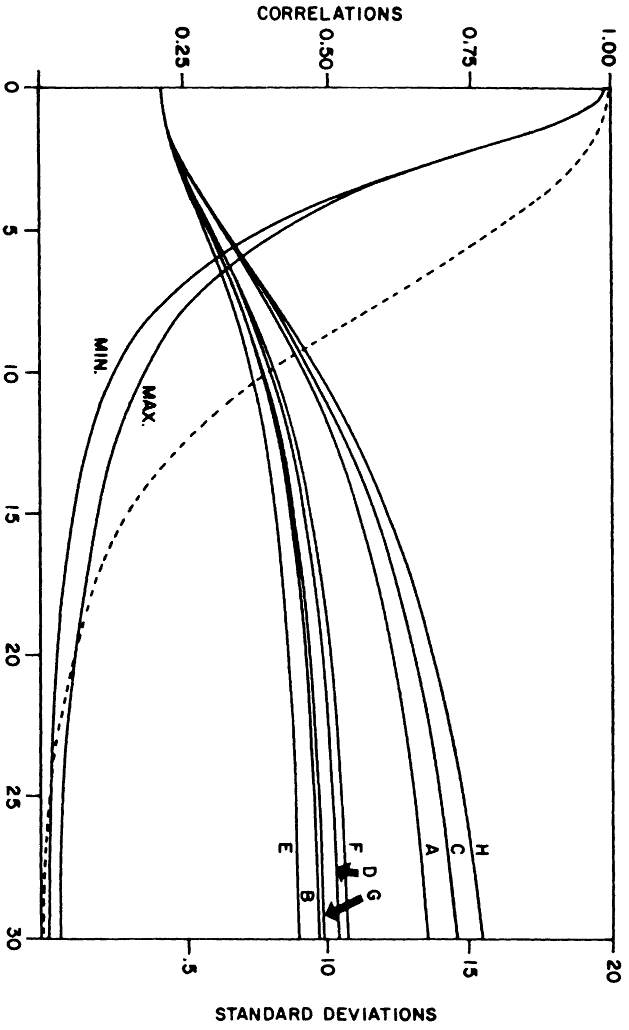

Figure 2: Standard Deviations of Treatment Effects as Function of S_e (ascending lines), Range of Correlations Among Treatment Effects as Function of S_e (descending lines), Likelihood of S_e (dotted line)

dividing points for even money bets on the results of new studies in the eight schools. However, the inferences conditional on maximum likelihood estimates of $S*$ would still underestimate the uncertainty in the estimates of μ_i because they would not reflect the uncertainty in the estimation of $S*$; consequently, these inferences might tend to underestimate the possibility of extreme effects.

The rationale for conditioning on one value of $S*$ is to simplify the summary of the study by avoiding the collection of conditional inferences generated by the possible values of $S*$ with their relative likelihoods. A more generally satisfactory way to simplify the summary of the study than conditioning on one value of S_* is to average the conditional inferences over the values of S_*, where the weights to attach to the conditional inferences are functions of the relative likelihoods of the values of S_*. This averaging procedure effectively considers S_* to be a random variable and thus requires the specification of a prior distribution for S_*.

V. A SUMMARY BAYESIAN ANALYSIS

Although we could use Figures 1 and 2 to summarize the effects in the eight studies, it is tempting to treat the likelihood of S_* depicted in the figures as the posterior distribution of S_* (corresponding to a uniform prior distribution on $S*$) and average answers over it. Although this process is easily done numerically, we illustrate a simple Monte Carlo method outlined in the Appendix.

First approximate the posterior distribution of S_* by considering many values of S_*, say $S_{*j} = j - 1/2$, $j = 1, 2, \ldots, 100$, with associated probability P_j, which is equal to expression (11) evaluated at S_{*j} divided by the sum of expression (11) from $j = 1$ to 100. This numerical approximation to the posterior distribution of S_* gives a value of 5 for the median value of S_*, 9 for the 75th percentile, 12 for the 90th percentile, 16 for the 95th percentile, and 25 for the 99th percentile. Draw a value of S_* from the posterior distribution (Step 1 of the Appendix provides details). Then draw μ_* from its posterior distribution given the drawn value of S_*; equations (6) and (7) are shown in Step 2 of the Appendix. Finally, independently draw $\mu_1, \ldots, \mu_8$ from their posterior distribution given $\mu*$, $S*$; equations (3)-(5) are shown in Step 3 of the Appendix. Step 4 of the Appendix will be discussed in Section VI.

The results of 200 simulations of school A's effect are shown in Display 1. A 95% interval for the effect in school A is (-2, 36). Table II summarizes the 200 simulated effect estimates for all eight schools. In one sense, the results in Table II are similar to the one common 95% interval ($8\underline{+}8$) of Section II in that the eight Bayesian 95% intervals are rather similar to each other and median-centered between 7 and 12. In a second sense, the results in Table II are quite different from the one common answer: the 95% intervals in Table II are larger than the one common interval and suggest substantially greater probabilities of effects larger than 16 points, especially in school A, and greater probabilities of negative effects, especially in school C. Of particular importance, the results in Display 1 are not similar to the separate answer for A: The Bayesian probability that the effect in school A is as large as 28 points is less than 10%, which is substantially less than the 50% probability based on school A's separate estimate.

Having performed the simulations outlined in the Appendix, it is easy to ask more complicated questions of this model. For example, what is the posterior distribution of the largest effect, $\max\{\mu_i\}$, from the eight schools? Display 2 provides the stem and leaf of 200 values from this posterior distribution and shows that only 17 (less than 10%) are larger than 28.4. Since Display 2 gives the marginal posterior distribution of the largest effect no matter which school it is in, and Display 1 gives the marginal posterior distribution of the effect for school A, Display 2 has larger values than Display 1.

VI. MONITORING THE BAYESIAN MODEL

The Bayesian summary presented in Section V is based on several model assumptions: (1) the normality of the estimates $\hat{\mu}_i$ given μ_i and V_i where the V_i are assumed known; (2) the exchangeability (i.e., i.i.d. nature) of the prior distribution of the μ_i; (3) the normality of the prior distribution

Display 1: Stem-and-leaf of 200 simulated values of treatment effects in school A.

```
  1     -1 | 1
  2    -0. | 8
  3      S | 7
  4      F | 4
  7      T | 322
  7     -0 |
 14      0 | 0001111
 27      T | 2222333333333
 46      F | 4444445555555555555
 62      S | 6666666667777777
 81     0. | 8888888888889999999
 31      1 | 0000000000000000011111111111111
 88      T | 22222222333333333
 71      F | 44444444445555
 59      S | 66666666677777
 45     1. | 8889999
 38      2 | 000111
 32      T | 2223333
 25      F | 44445
 20      S | 6677777
 13     2. | 9
 12      3 | 01
 10      T | 23
  8      F | 45
  6      S | 667
  3     3. | 9
  2      4 | 0
  1      T |
  1      F |
  1      S |
  1     4. | 9
```

TABLE II

Summary of 200 Simulated Treatment Effects for the Eight Schools

School	95% Interval	50% Interval	Median	50% Interval	95% Interval
A	-2	6	11	17	36
B	-6	4	8	12	19
C	-10	3	7	11	22
D	-7	4	7	13	21
E	-9	3	7	11	16
F	-8	2	7	11	20
G	-1	6	9	14	24
H	-3	4	8	13	24

Display 2: Stem-and-leaf of 200 values from the posterior distribution of the largest effect, $\mu_{(1)}$.

```
  1      T | 2
  2     -0 | 0
  5      0 | 011
  8      T | 233
 16      F | 4444555
 27      S | 66777777777
 46     0. | 8888888888889999999
 72      1 | 00000000001111111111111111
 96      T | 222222222222233333333333333
 16      F | 44444455555555555
 88      S | 66666666666677777
 71     1. | 888888889999999
 56      2 | 000000011111
 44      T | 22222333333
 32      F | 4445555
 25      S | 6667777
 18     2. | 8999
 14      3 | 111
 11      T | 22
  9      F | 4455
  5      S | 67
  3     3. | 9
  2      4 | 0
  1      T |
  1      F |
  1      S |
  1      4 | 9
```

of each μ_i given μ_*, S_* and (4) the uniformity of the prior (marginal) distribution of μ_*, S_*. The first modelling assumption will not be questioned here for the reasons given in Section III.

The second modelling assumption deserves commentary. The real-world interpretation of this mathematical assumption of exchangeability of the μ_i is that before seeing the results of the experiments, there is no desire to include in the model features such as (a) the effect in school A is probably larger than in school B, or (b) the effects in schools A and B are more similar than in schools A and C. In other words, the exchangeability assumption means that we will let the data tell us about the relative ordering and similarity of effects in the schools. Such an a priori stance seems reasonable when the results of eight parallel experiments are being scientifically summarized for general presentation; of course, there might exist generally accepted information about the programs or the schools to make us formulate a non-exchangeable prior distribution (e.g., schools B and C have similar students and schools A,D,E,F,G,H have similar students. Based on a liberal interpretation of a mathematical result in de Finetti 1963, the exchangeability assumption in our problem implies that the prior distribution of the μ_i can be chosen to be i.i.d. given some hyperparameters (e.g., μ_*, V_*) that also have a prior distribution.

The third and fourth modelling assumptions are harder to justify a priori than the first two. Why should school effects be normally distributed rather than say, Cauchy distributed (or even asymmetrically distributed), and why should the location and scale parameters of this prior distribution (e.g., μ_* and S_*) be uniformly distributed? Mathematical tractability is one reason for the choice of models, but if the family of Bayesian models is inappropriate, Bayesian answers can be quite misleading.

The method proposed here for checking the adequacy of the models is called "phenomenological Bayesian monitoring". The idea is to monitor the entire model being used to make sure that it could generate the observed data; the spirit is similar to that of significance testing but is Bayesian in that all inferences are conditional on the observed data. It is also quite similar to the prior predictive check tests described by Box (1980). Our methodology, however, requires only that the posterior distribution is consistent with observed data whereas Box's approach monitors the entire prior specification.

Specifically, our method is as follows. Under the model for the data, there exists a distribution of the model parameters conditional on the observed data. This posterior distribution can be used to generate a posterior predictive distribution for the data in a new study replicating the study from which we have data. The posterior predictive distribution of the data from the new study is compared to (a) the actual data from the study and (b) scientific judgment about plausible values of such data. If the model is an appropriate one, the hypothetical data generated under the model will not be in conflict with the actual data or with scientifically plausible values, but will be typical of them. Unless there exists very strong prior knowledge, the use of models that are contradicted by observed data is to be avoided. The adjective "phenomenological" is used to describe this monitoring of Bayesian models because of the emphasis on comparing observed quantities to predictions about similar observable quantities. That is, the focus when checking the model is on phenomena, not unobservable parameters.

Of course, for any one data set there are literally an infinite number of models that can generate data consistent with the observed data, and so scientific judgment and practical computational tractability are also important criteria for choosing models. In some cases, an important component of a data analysis may be to show how predictions of new data, especially extrapolations, are sensitive to models all of which are consistent with the observed data; it is hopeless, however, to try to display such sensitivity to all models that are consistent with the observed data.

We generated 200 simulated studies of eight experiments under the Bayesian model used in Section V. The simulation method is outlined in Step 4 of the Appendix and simply extends the Monte Carlo presented in Section V by drawing hypothetical values of observed effects, $\tilde{\mu}_1,\ldots,\tilde{\mu}_8$, from their posterior predictive distribution. For any exchangeable prior distribution on the μ_i, the sufficient statistics for estimating this prior distribution are the eight vectors:

$(\tilde{\mu}_{(1)}, V_{(1)}^{\frac{1}{2}}), \ldots, (\tilde{\mu}_{(8)}, V_{(8)}^{\frac{1}{2}})$ where $(\tilde{\mu}_{(1)}, V_{(1)}^{\frac{1}{2}})$ is the value of $(\tilde{\mu}_i, V_i^{\frac{1}{2}})$ for the largest estimated effect, $(\tilde{\mu}_{(2)}, V_{(2)}^{\frac{1}{2}})$ is the value of $(\tilde{\mu}_i, V_i^{\frac{1}{2}})$ for the second largest estimated effect, and so on.

Figure 3 presents the joint distribution of the largest estimated effect, $\mu_{(1)}$ and its associated standard error, $V_{(1)}^{\frac{1}{2}}$. Since each school has a unique value of $\underset{\sim}{V_i}$ which is assumed known, this plot is basically a plot of $\mu_{(1)}$ within each school, where the schools are identified by their V_i. From this figure, in 41 of 200 instances, the largest effect occurred in school A and of these 41, in 28 instances the estimated effect was larger than the observed value of 28.4. The corresponding numbers for the other effects are presented in Table III. This summary suggests that the model generates observed values similar to the observed values in our study; that is, the observed values are typical of the estimated effects generated by the model.

Not only are the observed values of the data consistent with the values generated under the model, but in addition, the model-generated values are all plausible outcomes of experiments on coaching. The smallest estimated effect generated was -42 and the largest estimated effect generated was 66; because both values are possible for estimated effect sizes from studies of the SAT, all estimated values generated by the model are possible outcomes. Hence, the model seems to generate data that are consistent with observed values and scientific judgment, and so the conclusions in Section V based on this model are respectable.

Of course, there are many functions of the simulated $(\tilde{\mu}_{(i)}, V_{(i)}^{\frac{1}{2}})$ that we could examine such as their raw values or the gaps $(\tilde{\mu}_{(1)} - \tilde{\mu}_{(2)}), \ldots, (\tilde{\mu}_{(7)} - \tilde{\mu}_{(8)})$. If we had a particular nonnormal prior distribution in mind for the μ_i, the sufficient statistics under that model would guide our choice of which functions of $(\tilde{\mu}_{(i)}, V_{(i)}^{\frac{1}{2}})$ to examine. Often

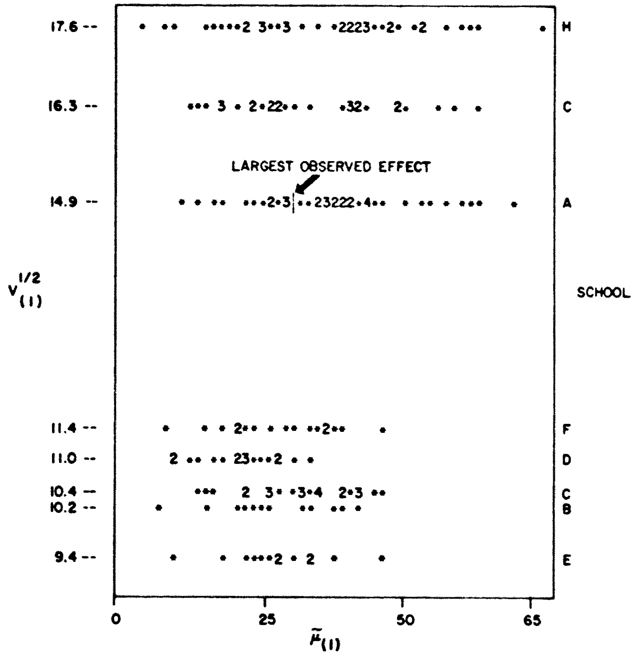

Figure 3: Joint Distribution of Largest Estimated Effect $\tilde{\mu}_{(1)}$, and its Associated Standard Error, $V_{(1)}^{1/2}$: 200 Simulated Values

TABLE III

Summary of 200 Simulations of the Estimated Effects for the Eight Schools

Number of Times That	i = 1	2	3	4	5	6	7	8
ith largest estimated effect occured in the school with ith largest observed effect	41	25	18	19	24	23	30	27
and								
was larger than ith largest observed effect	28	11	10	10	8	16	13	5

in practice, however, it is easier to obtain diagnostically useful displays directly from intuitively interesting statistics than to derive analytically the sufficient statistics from interesting alternative models. This is a major thrust of exploratory data analysis (Tukey, 1980).

Two final points. First, because we act as if the model is acceptable does not mean that there are no other models that fit the data just as well; there are an infinite number of such models and conclusions from a data analysis are always somewhat tentative and dependent upon assumptions. Second, if we found out that the model did not fit, we would have been obliged to search for a new model that did fit; a data analysis is rarely, if ever, complete with simply a rejection of some model.

REFERENCES

Alderman, D. & Powers, D. The effects of special preparation on SAT-Verbal scores. Research Report 79-1. Princeton, N.J.: Educational Testing Service, 1979.

Box, G.E.P. Sampling and Bayes' inference in scientific modelling and Robustness. Journal of the Royal Statistical Society, Series A, 1980, 143, 383-430.

Box, G.E.P. & Tiao, G. Bayesian inference on statistical analysis. Reading, MA.: Addison-Wesley, 1973.

de Finetti, B. Foresight: its logical laws, its subjective sources. In H.E. Kyburg & H.E. Smokler, (eds.), Studies in subjective probability. New York: Wiley, 1963.

Dempster, A.P., Rubin, D.B. & Tsutakawa, R.K. Estimation in covariance components models. Journal of the American Statistical Association, 1981, 76, 341-353.

Efron, B. & Morris, C. Stein's paradox in statistics. Scientific American, 1977, 236(5), 119-127.

James, W. & Stein, C. Estimation with quadratic loss. Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 1, Berkeley and Los Angeles: University of California Press, 1961.

Lindley, D.V. & Smith, A.F.M. Bayes estimation for the linear model (with discussion). *Journal of the Royal Statistical Society*, *Series B*, 1972, *34*, 1-41.

Lord, F.M. & Novick, M.R. *Statistical theories of mental test scores*. Reading, MA.: Addison-Wesley, 1968.

Novick, M.R., Jackson, P.H., Thayer, D.T. & Cole, N.S. Estimating multiple regression in m groups: a cross-validation study. *The British Journal of Mathematical and Statistical Psychology*, 1972, *25*, 33-50.

Rubin, D.B. Using empirical Bayes techniques in the law school validity studies. *The Journal of the American Statistical Association*, 1980, *75*(372), 801-827.

Shigemasu, K. Development and validation of a simplified m-group regression model. *Journal of Educational Statistics*, 1976, *1*(2), 157-180.

Tukey, J.W. *Exploratory data analysis*. Reading, MA.: Addison-Wesley, 1978.

Wang, M. et al. A Bayesian data analysis system for the evaluation of social programs. *The Journal of the American Statistical Association*, 1977, *72*, 711-722.

AUTHOR

Donald B. Rubin. Address: Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541. Title: Senior Statistical Research Advisor. Degrees: A.B. Princeton University, M.S. Harvard University, Ph.D. Harvard University. Specializations: Causal inference in randomized and nonrandomized studies; sample survey methods; missing data problems; Bayes and empirical Bayes techniques.

APPENDIX

Simulating the posterior distribution of (1) the effects $(\mu_1,\ldots,\mu_K)$ and (2) the estimated effects, $(\tilde{\mu}_1,\ldots,\tilde{\mu}_K)$, in new experiments replicating the observed experiment.

For each replication, perform Steps 1-4 where $(\hat{\mu}_1,\ldots,\hat{\mu}_K)$, $(V_1,\ldots,V_K)$ and K are given; $\lambda_i = S_*{}^2/(V_i + S_*{}^2)$; and Z refers to an independent normal deviate (a new draw each time Z appears).

Step 1: Draw S_* from its posterior distribution:

$S_* = J-\frac{1}{2}$ where J is the smallest positive integer such that $P(J) \geq u$, where

u = uniform random number on (0,1) and

$$P(J) = \sum_{j=1}^{J} L(j-\tfrac{1}{2}) / \sum_{j=1}^{100} L(j-\tfrac{1}{2}),$$

$$L(X) = \left[\frac{\prod_{i=1}^{K} \lambda_i}{X^{2K-2} \sum_{i=1}^{K} \lambda_i}\right]^{\frac{1}{2}} \exp\left\{-\frac{1}{2X^2}\left[\sum_{i=1}^{K} \lambda_i \hat{\mu}_i^2 - \frac{\left(\sum_1^K \hat{\mu}_i \lambda_i\right)^2}{\sum_1^K \lambda_i}\right]\right\}$$

Step 2: Draw μ_* from its posterior distribution given S_*:

$$\mu_* = \sum_{i=1}^{K} \hat{\mu}_i \lambda_i / \sum_{i=1}^{K} \lambda_i + \left(S_*^2 / \sum_{i=1}^{K} \lambda_i\right)^{\frac{1}{2}} \times Z$$

Step 3: Draw $(\mu_1,\ldots,\mu_K)$ from its posterior distribution given (μ_*,S_*):

$$\mu_i = \lambda_i\hat{\mu}_i + (1-\lambda_i)\mu_* + (\lambda_i V_i)^{\frac{1}{2}} \text{ x } Z_i, \quad i=1,\ldots,K.$$

Then $(\mu_1,\ldots,\mu_K)$ simulates the posterior distribution of the effects in the K experiments.

Step 4: Draw $(\tilde{\mu}_1,\ldots,\tilde{\mu}_K)$ from its posterior distribution given $(\mu_1,\ldots,\mu_K,\mu_*,S\ )$:

$$\tilde{\mu}_i = \mu_i + V_i^{\frac{1}{2}} \text{ x } Z_i, \quad i=1,\ldots,K.$$

Then $(\tilde{\mu}_1,\ldots,\tilde{\mu}_K)$ simulates the estimated effects in K experiments replicating the K observed experiments.