Abstract Title Page
Not included in page count.

Title:

Distortions in Distributions of Impact Estimates in Multi-Site Trials: The Central Limit Theorem is Not Your Friend.

Authors and Affiliations:

Henry May
Director and Associate Professor
Center for Research in Education and Social Policy (CRESP)
University of Delaware
hmay@udel.edu

Abstract Body

Limit 4 pages single-spaced.

Problem / Background / Context:

Interest in variation in program impacts—How big is it? What might explain it?—has inspired recent work on the analysis of data from multi-site experiments, much of which has been presented at SREE (Bloom, 2012; Bloom, Porter, & Weiss, 2013; May & D'Agostino, 2012; Raudenbush, Bloom, & Reardon, 2014; Weiland & Bloom, 2014). One critical aspect of this problem involves the use of random or fixed effect estimates to visualize the distribution of impact estimates across a sample of sites. Unfortunately, unless the sample sizes within sites are enormous (e.g., >1,000), the shape of the distribution of estimates is dominated by the normal distribution as a result of imperfect reliability in site-specific estimates. The Central Limit Theorem explains why this occurs and why the problem is unavoidable.

Purpose / Objective / Research Question / Focus of Research:

The purpose of this paper is to demonstrate the scope and severity of this problem, and to provide cautionary guidance regarding the interpretation of distributions of site-specific impact estimates in multi-site experiments. Furthermore, the paper explores methods for recovering the distribution of true site-specific impacts when that distribution is non-normal.

Improvement Initiative / Intervention / Program / Practice:

For illustrative purposes, data from the multi-site randomized trial under the Reading Recovery Investing in Innovation (i3) Scale-Up is used in addition to generalized Monte Carlo simulation. Reading Recovery is an early reading intervention for first-graders who are reading substantially below grade level. Students in Reading Recovery participate in one-to-one lessons with a reading specialist for 30 minutes each day for up to 20 weeks.

Setting:

The Reading Recovery i3 multi-site RCT is being conducted in elementary schools participating in the $55 million i3 Scale-Up of Reading Recovery. At the start of the 2012-13 school year, the grant's third year, 2,085 teachers have been trained in Reading Recovery or are in training, with support from i3 funds. The majority of these teachers are working in high-need schools such as buildings that are persistently underperforming or that enroll high numbers of English language learners.

Population / Participants / Subjects:

The data used here from the Reading Recovery i3 multi-site RCT includes 2,296 students from 380 schools across the nation. Up to eight students in each school were matched into pairs based on pretest text reading level and English language learner status. One student in each matched pair was randomly assigned to participate in the intervention during the first half of the school year, while the other student in that pair was assigned to the control group (i.e., eligible to participate in Reading Recovery during the second half of the school year). Post-intervention

outcomes were measured for both students in a matched pair when the treatment student completed or otherwise left the intervention.

Research Design:

The research design for this study involves two components: (1) Data from the Reading Recovery multi-site RCT is analyzed to determine overall program impacts, to estimate variation in impacts across sites, and to visualize the distribution of site-specific impact estimates; (2) Monte Carlo simulation is used to show how different non-normal distributions of true site-specific effects are largely concealed in visualizations based on both fixed effects and random effects analyses, with the latter estimated using both maximum likelihood and Bayesian simulation.

Data Collection and Analysis:

Data from 1,148 treatment students and 1,148 control students from 380 schools participating in the Reading Recovery multi-site randomized trial were analyzed to produce an overall impact estimate, estimates of site-level variation in impacts, and visualizations of the distribution of site-specific impacts. Differences in performance of the treatment and control students was estimated using a three-level hierarchical linear model (HLM) (Raudenbush & Bryk, 2002), with students nested within blocks (i.e., matched pairs) and blocks nested within participating schools. The HLM includes pretest scores as a covariate, along with random effects for blocks, a random effect for overall school performance (i.e., random school intercepts), and a random effect for the impact of Reading Recovery (i.e., random treatment effects across schools). The primary impact analyses utilizes the reading words and reading comprehension subscales from the Iowa Tests of Basic Skills (ITBS) as the posttest outcome measure and scores from the Observation Survey of Early Literacy Achievement (OS) as the pretest covariate.

Monte Carlo analyses were conducted in order to produce estimates of site-level variation in impacts and visualizations of the distribution of site-specific impacts, with known non-normal distributions of true site-specific impacts. The non-normal distributions were uniform, lognormal, and negative lognormal. A mean of .50 and a standard deviation of .50 was used for each non-normal distribution of true treatment effects (i.e., consistent with the Reading Recovery mixed model estimates of treatment effects and variation). Each simulated dataset was subjected to three types of analytic models: (1) fixed effects via OLS, (2) empirical Bayes estimates from HLM via maximum likelihood, and (3) posterior distributions of random effects from HLM using Bayesian simulation.

Findings / Outcomes:

HLM analyses revealed overall effects of Reading Recovery in this multi-site study are large and positive (i.e., Glass' Δ between .36 and .68) with substantial variation in impact estimates across sites (i.e., $\sqrt{\tau} \cong .50$ where τ is the variance component for site-specific random treatment effects). The distributions of site-specific treatment effect estimates appeared remarkably normal, with a few positive outliers.

Results from the Monte Carlo simulations suggests that random effects HLM models do a reasonable job of producing an unbiased estimate of treatment effect variation (i.e., the variance of the true impacts $\sqrt{\tau} \cong .50$) regardless of the shape of the true distribution of impact estimates. However, none of the analytic approaches (i.e., fixed effects via OLS, random effects via HLM, or random effects via Bayesian simulation) were able to recover the original distributions of true site-specific impacts. In all cases, the distribution of estimates appeared remarkably normal, in stark contrast to the non-normal distributions of the true site-specific effects. It was only when within-site sample sizes were increased dramatically (e.g., $n_j = 200$) that the non-normal distributions began to appear clearly in the distributions of fixed and random effects impact estimates.

Conclusions:

The reasons for these results are simple. The Central Limit Theorem states that each site-specific estimate is distributed around its corresponding true estimate as a normal distribution with a mean of zero (i.e., the estimate is unbiased) and a variance of $4\sigma^2/n_j$ (i.e., the standard error of the mean difference, assuming a balanced design in each site). The reliability of these estimates is simply the ratio of true site-specific impact variability to the total variability in site-specific impact estimates (i.e., $\tau/[\tau + (4\sigma^2/n_j)]$). In the Reading Recovery study, the reliability of the Empirical Bayes estimates of site-specific impacts was approximately .40, while in the simulations reliability was .38 with n_j =10 and .92 with n_j=200. When the reliability of the estimates is relatively low (e.g., r <.50), the shape of the distribution of estimates is dominated by the normal errors. Only when the reliability is high (e.g., r >.90) does the distribution of true site-specific impacts begin to dominate the shape of the distribution of impact estimates.

Given that most RCTs are powered only to detect overall program effects, and that power curves for multi-site RCTs typically level off with relatively low numbers of subjects per site, it is reasonable to expect that within-site sample sizes n_j will be relatively small (e.g., <30) and reliabilities relatively low (r <.50) in the context of a multi-site experiment involving a large enough number of sites to justify visualization of the distribution of site-specific estimates (e.g., J > 30).

Future work on this paper between now and the fall SREE conference will include methods intended to separate the distribution of true site-specific impacts from the distribution of normal sampling errors. Results thus far suggest that finite mixture models (FMM) may hold promise, although the typical parameterization of FMM assumes that each data point is observed from a distinct distribution. In this case, each data point is equal to the sum of points from two distinct distributions; however, the knowledge that one distribution is normal with a known mean and variance may allow the recovery of the unknown distribution of true treatment effects.

Appendices

Not included in page count.

Appendix A. References

Bloom, H. (2012). *Impact Variation: How Do You Know It When You See It?* Keynote address given at the meeting of the Society for Research in Educational Effectiveness (SREE), March, 2012.

Bloom, H., Porter, K., & Weiss, M. (2013). *Estimating Cross-Site Impact Variation in the Presence of Heteroscedasticity*. Paper presented at the meeting of the Society for Research in Educational Effectiveness (SREE), September 2013.

May, H. & D'Agostino, J. (2012). *Exploring Treatment Variation in the Scale-Up of Reading Recovery*. Paper presented at the meeting of the Society for Research in Educational Effectiveness (SREE), March 2012.

Raudenbush, S., Bloom, H., & Reardon, S. (2014). *Using Cross-Site Variation in Program Effects to Study What Works for Whom, Under What Conditions, and Why*. Workshop conducted at the meeting of the Society for Research in Educational Effectiveness (SREE), March 2014.

Weiland, C. & Bloom, H. (2014). *To What Extent Do Head Start's Effects on Children's Language, Literacy, Mathematics, and Socio-Emotional Skills Vary Across Individuals, Subgroups, and Centers?* Paper presented at the meeting of the Society for Research in Educational Effectiveness (SREE), March 2014.

Appendix B. Tables and Figures

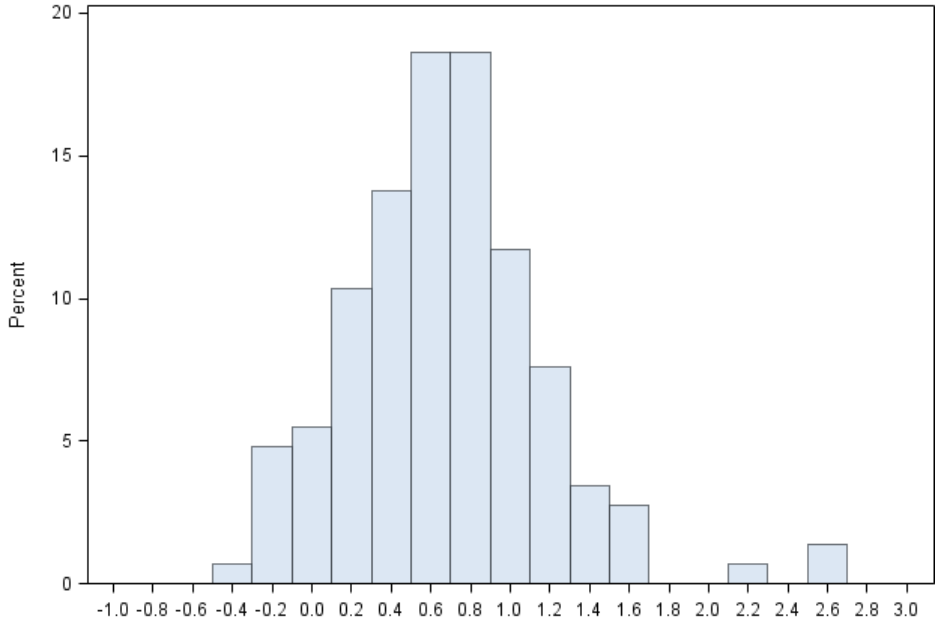

Figure 1. Adjusted Empirical Bayes Site-Specific Glass' D Effect Size Estimates from the 2011-12 Reading Recovery Multi-Site Randomized Trial

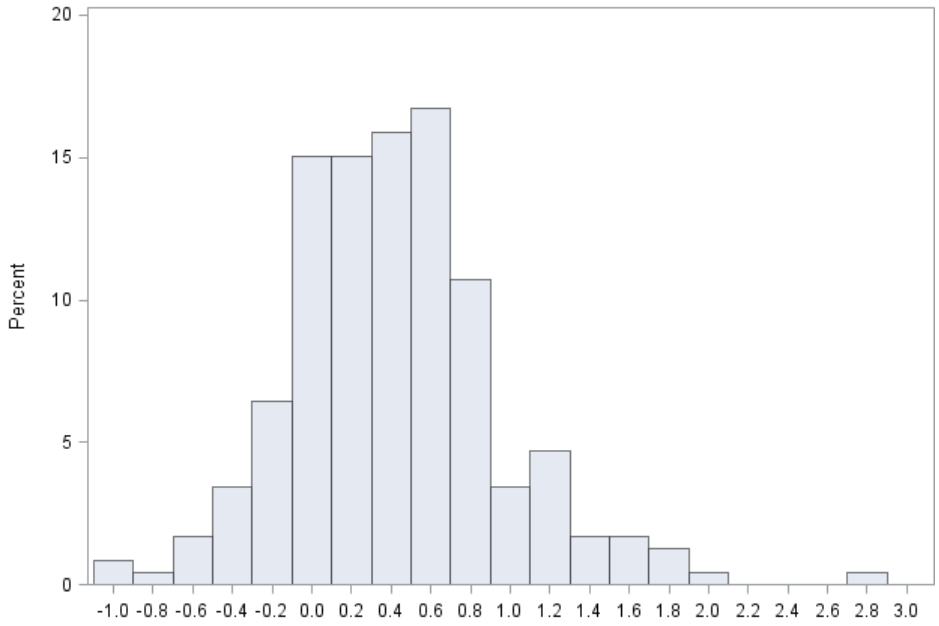

Figure 2. Adjusted Empirical Bayes Site-Specific Glass' D Effect Size Estimates from the 2012-13 Reading Recovery Multi-Site Randomized Trial

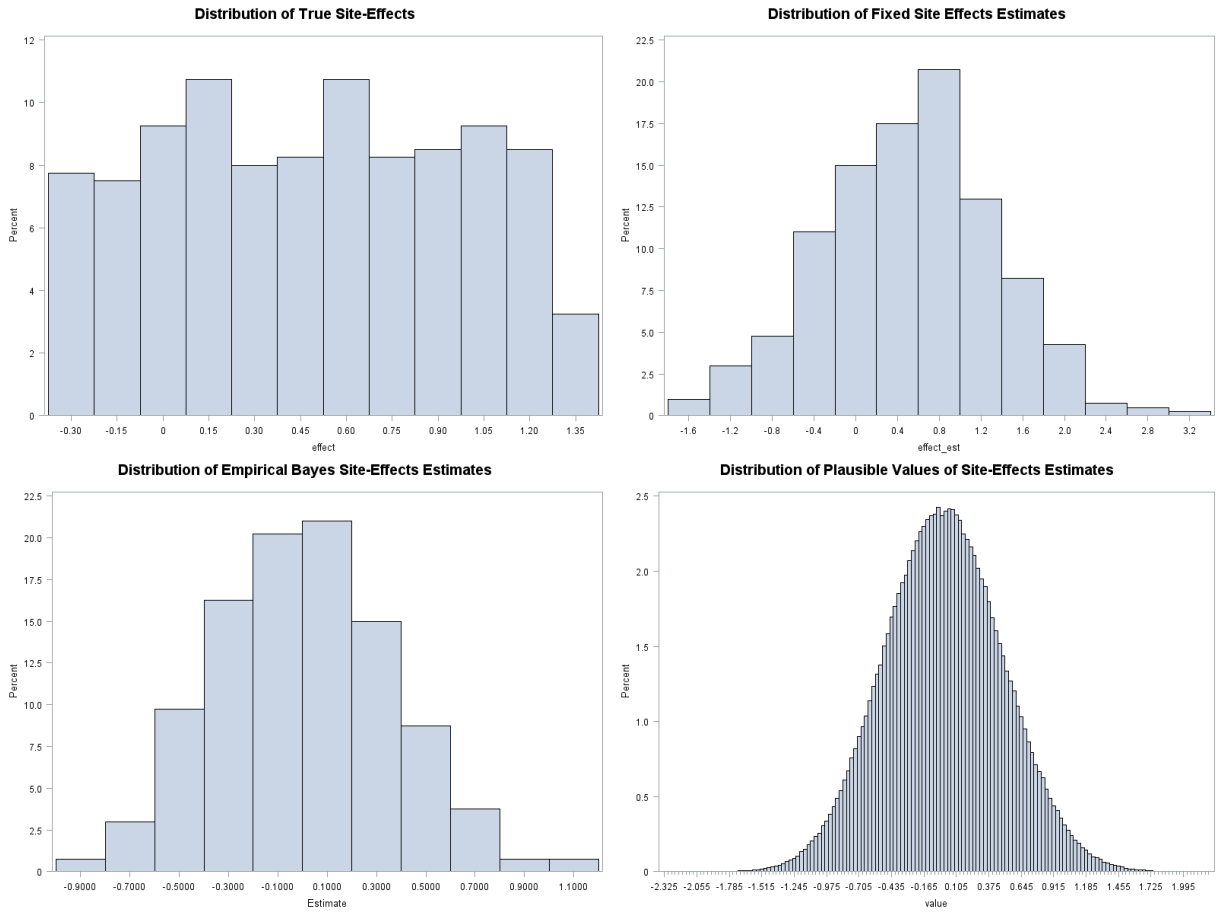

Figure 3. True and Estimated Distributions of Site-Specific Treatment Effects when True Effects are Drawn from a Uniform Distribution (400 sites, 5 treatment and 5 control subjects per site, mean δ=.50, SD_δ=.50)

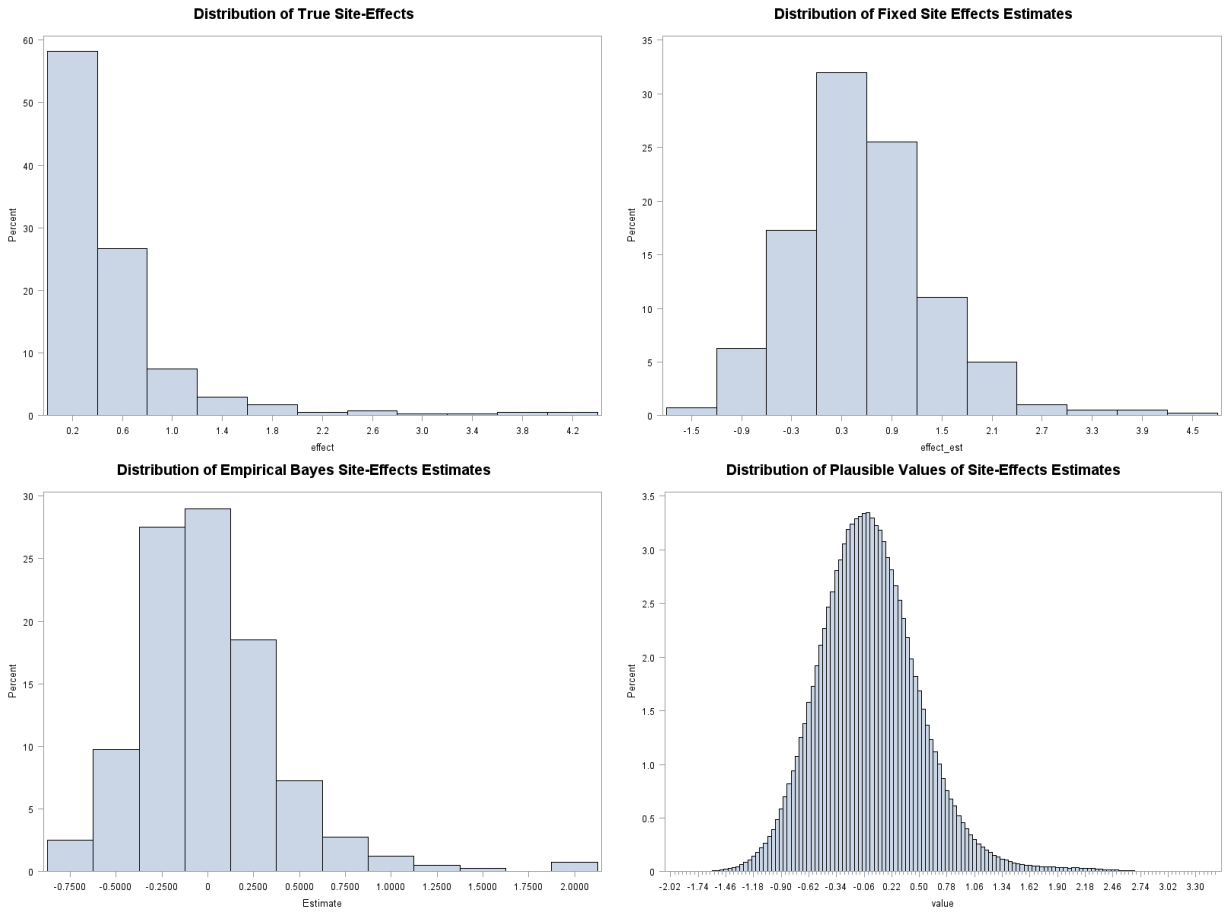

Figure 4. True and Estimated Distributions of Site-Specific Treatment Effects when True Effects are Drawn from a Lognormal Distribution (400 sites, 5 treatment and 5 control subjects per site, mean δ=.50, SD_δ=.50)

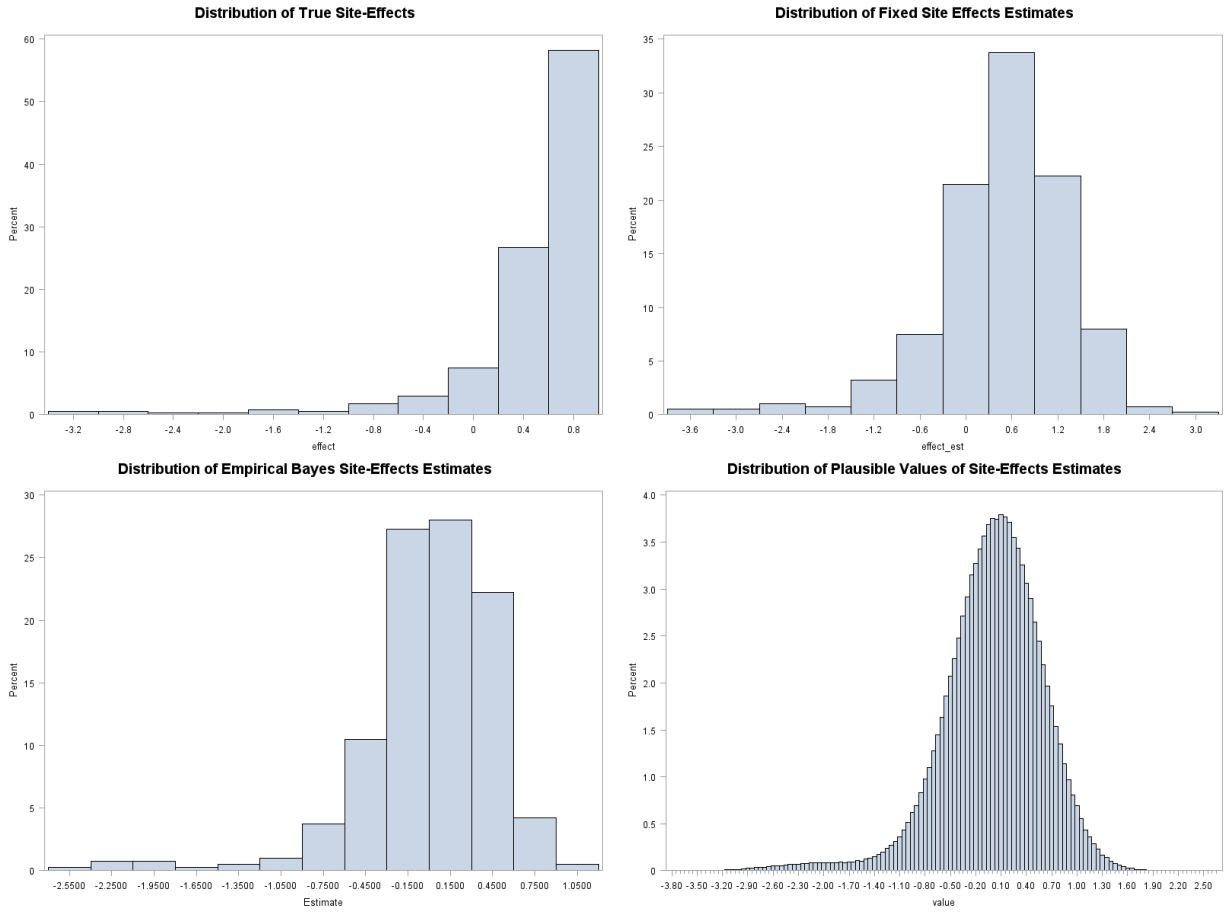

Figure 5. True and Estimated Distributions of Site-Specific Treatment Effects when True Effects are Drawn from a Negative Lognormal Distribution (500 sites, 5 treatment and 5 control subjects per site, mean δ=.40, SD_δ=.10)

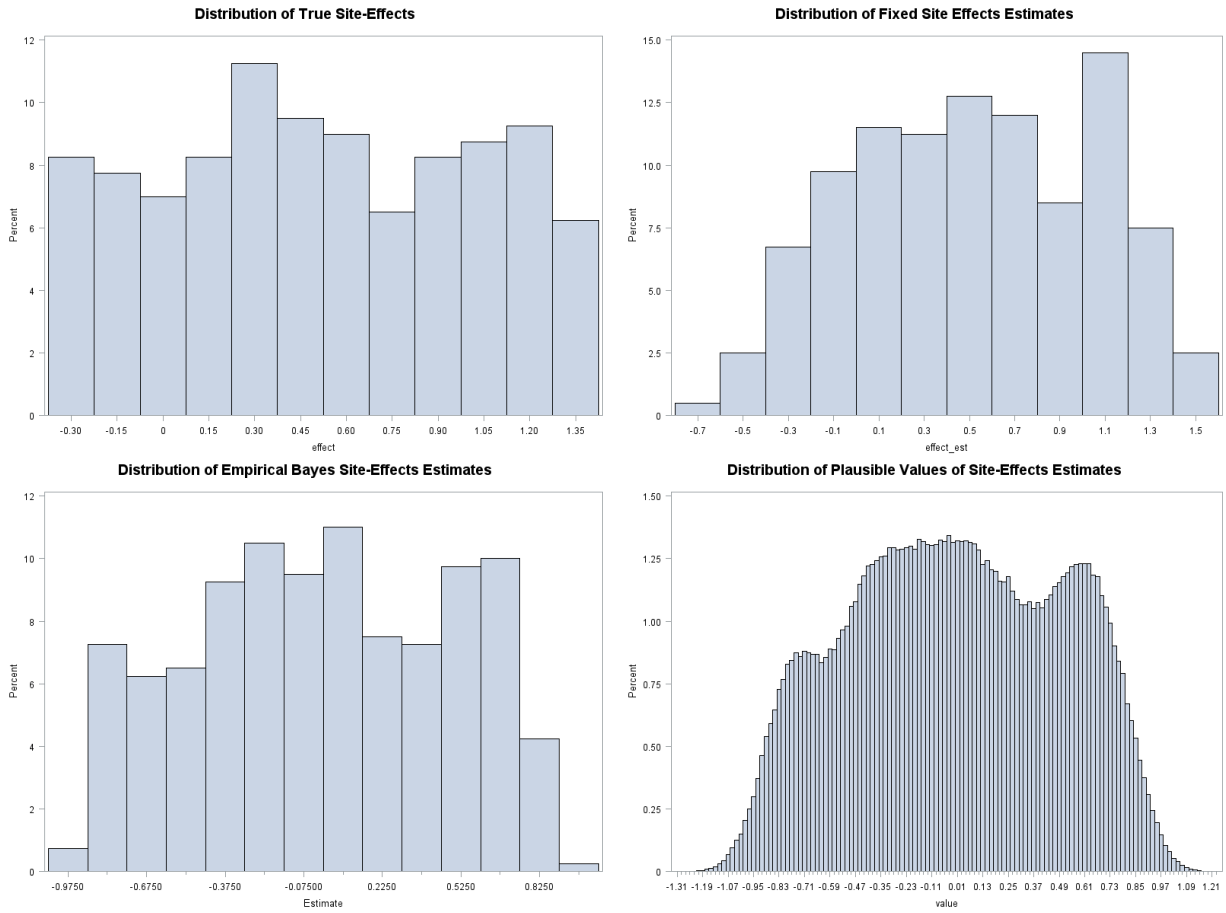

Figure 7. True and Estimated Distributions of Site-Specific Treatment Effects when True Effects are Drawn from a Uniform Distribution (400 sites, 100 treatment and 100 control subjects per site, mean δ=.50, SD_δ=.50)

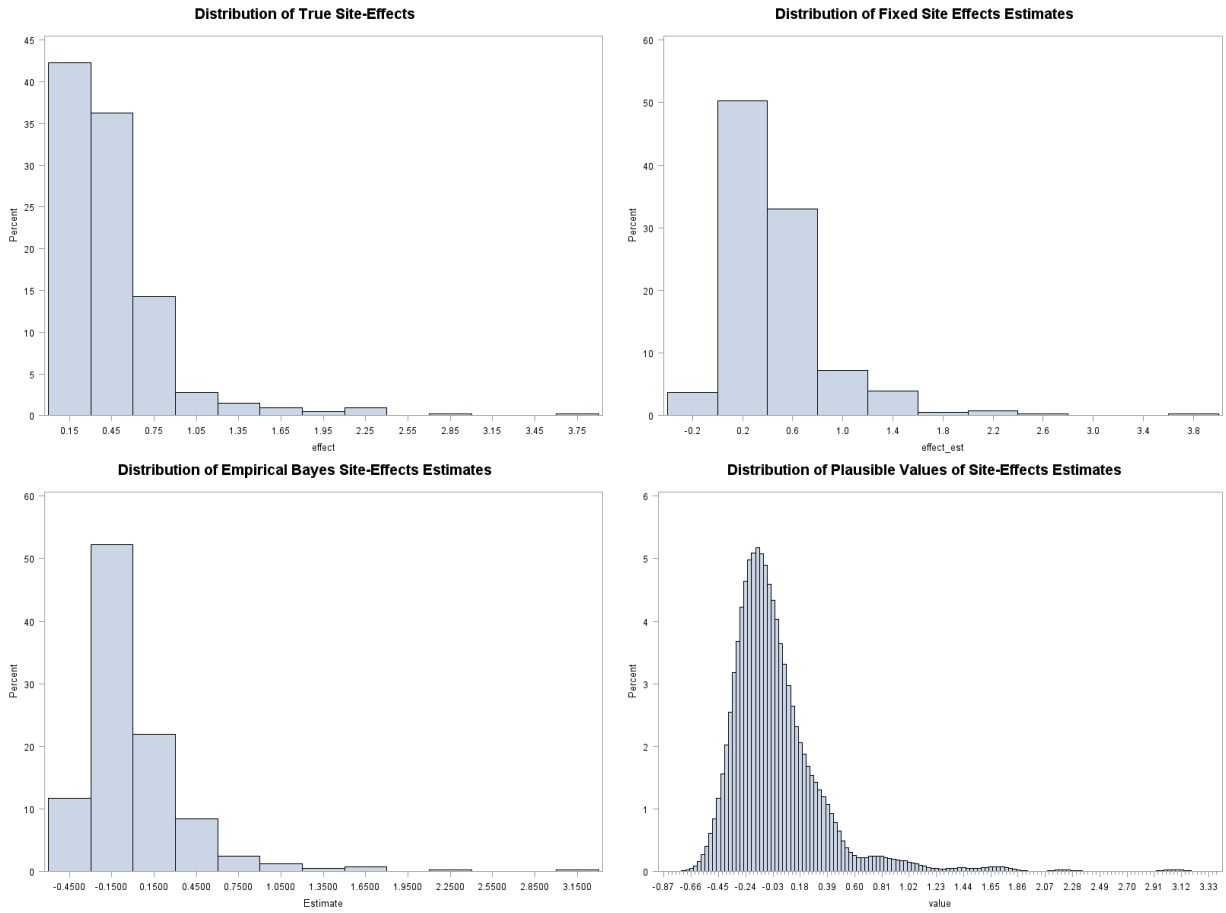

Figure 8. True and Estimated Distributions of Site-Specific Treatment Effects when True Effects are Drawn from a Lognormal Distribution (400 sites, 100 treatment and 100 control subjects per site, mean δ=.50, SD_{δ}=.50)

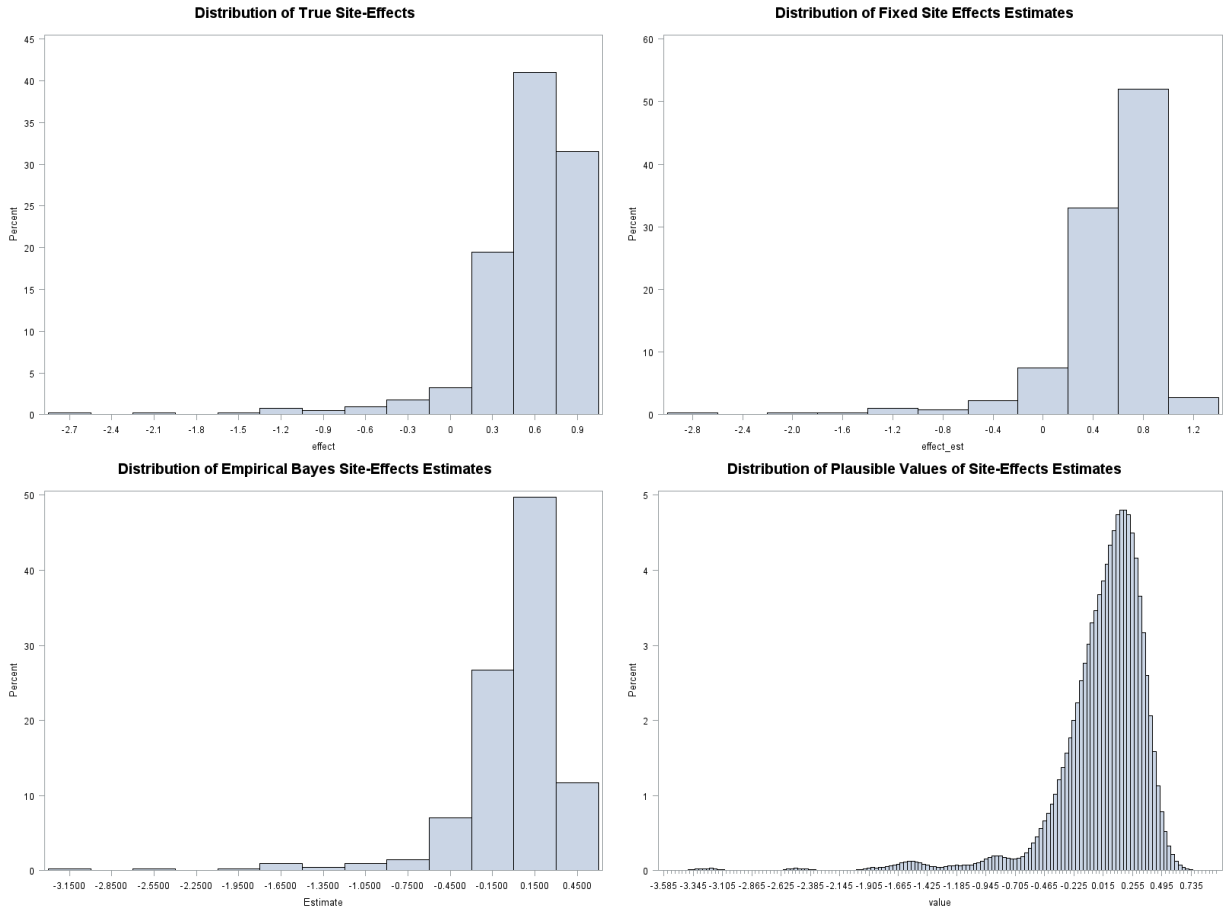

Figure 9. True and Estimated Distributions of Site-Specific Treatment Effects when True Effects are Drawn from a Negative Lognormal Distribution (400 sites, 100 treatment and 100 control subjects per site, mean δ=.50, SD_δ=.50)