Bayesian Partial Pooling to Improve Inference Across A/B Tests in EDM

Adam C Sales
University of Texas at Austin
536C George I. Sánchez Building
Austin, TX 78705
asales@utexas.edu

Thanaporn Patikorn
Worcester Polytechnic Institute 100 Institute Rd
Worcester, MA 01609
tpatikorn@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute 100 Institute Rd
Worcester, MA 01609
nth@wpi.edu

ABSTRACT

This paper will explain how analyzing experiments as a group can improve estimation and inference of causal effects–even when the experiments are testing unrelated treatments. The method, composed of ideas from meta-analysis, shrinkage estimators, and Bayesian hierarchical modeling, is particularly relevant in studies of educational technology. Analyzing experiments as a group–"partially pooling" their respective datasets–increases overall accuracy and avoids issues of multiple comparisons, while incurring small bias. The paper will explain how the method works, demonstrate it on a set of randomized experiments run within the ASSISTments platform, and illustrate its properties in a simulation study.

1. INTRODUCTION

Using educational technology to conduct many experiments, as in the ASSISTments TestBed [7], allows education researchers to rigorously answer many causal questions and test many hypotheses independently. Perhaps more surprisingly, the various experiments can help *each other*. Effect estimates that partially pool data across experiments—even those that are testing very different interventions—are often more precise and accurate, and less error-prone, than estimates based on the experiments individually.

This poster will illustrate a Bayesian approach to analyzing several experiments simultaneously. (By "Bayesian," here, we mean merely that the goal of the approach is a posterior distribution for treatment effects.) The method combines ideas from [8] and [1] on shrinkage, from [12] on Bayesian partial pooling to examine treatment effect heterogeneity, and [6] on multiple comparisons. The paper's main contributions will be to introduce these ideas to an EDM audience—where, due to proliferation of online experiments, they are particularly applicable—and to illustrate their potential.

Previously, [10] combined data across experiments to improve covariance adjustment; that method is orthogonal, and perhaps complementary, to ours, which does not use covariates. [3], [9], and many others have used multilevel, hierarchical Bayesian modeling to analyze intelligent tutor data, but not in the context of experiments.

After describing and explaining the method (Section 2), we will illustrate it in an analysis of a dataset comprised of 22 parallel experiments run inside ASSISTments [13] (Section 3) and in a simulation study (Section 4). We will show that partially pooling data from across experiments increases precision while lowering type-I error rates, decreases the width of confidence intervals while improving their coverage, and substantially reduces the incidence of drawing incorrect conclusions from experimental data.

2. SHRINKAGE, PARTIAL POOLING, AND REGRESSION TO THE MEAN

Unbiased estimates $\hat{d}^{np}$ of effect sizes d from randomized A/B tests are noisy—a different estimate would have resulted had the treatment been randomized differently. The standard error of a particular effect size estimate, $\sigma_i = SD(\hat{d}_i^{np}|d_i)$, depends on a number of factors, most principally the sample size n_i, but in practice it is never zero. Similarly, among a group of K experiments, the true effect sizes d_i, $i = 1, ..., K$, (presumably) vary as well—$var(d) = \tau$, say. Considered together, the variance of a group of effect size estimates is the sum of both components: the variance of the true effects plus the average of the (squared) standard errors of the individual estimates:

$$var(\hat{d}^{np}) = \tau^2 + \mathbb{E}[\boldsymbol{\sigma}^{\mathbf{2}}]$$

In other words, the distribution of effect size *estimates* is wider than the distribution of true effect sizes. Therefore, the largest effect size estimates $\hat{d}^{np}$ typically *over*estimate their respective true effects d, and that the smallest effect size estimates typically *under*estimate their true effects. This is an example of regression to the mean [4] (also see [16]).

The implication for estimating effects can be startling. When A/B tests are analyzed independently, the best estimate for the true effect size d_i in experiment i is $\hat{d}_i^{np}$. However, when the K experiments are considered as a group, $\hat{d}^{np}$ is inadmissible. A better estimate, $\hat{d}^{pp}$, corrects for the fact that the extreme estimates are probably too extreme, and shrinks

them toward the overall mean effect size μ [2]:

$$\hat{d}_i^{pp} = \mu + c_i \left(\hat{d}_i^{np} - \mu \right) \quad (1)$$

where c_i is a "shrinkage coefficient" between 0 and 1. When $c_i = 1$, $\hat{d}_i^{pp} = \hat{d}_i^{np}$; when $c_i = 0$, $\hat{d}_i^{pp} = \mu$.

Another term for this procedure is "partial pooling" [5]. The overall mean treatment effect, μ, can be estimated by completely pooling the data across all K experiments. In contrast, individualized estimates $\hat{d}^{np}$ result if data from different A/B tests are not pooled at all—$\hat{d}^{np}$ is a "no pooling" estimate. The optimal estimate $\hat{d}^{pp}$ combines the the no-pooling estimate $\hat{d}^{np}$ with a complete-pooling estimate of μ—hence, partial pooling.

In general, the size of the shrinkage coefficient c_i, which regulates the extent of the partial pooling, depends both on the standard deviation of the true effects, τ, and σ_i, the standard error of $\hat{d}_i^{np}$. When τ is large, the experiments differ widely from each other, so the overall mean effect μ tells us little about the individual effects d. When σ_i is large, then $\hat{d}_i^{np}$ is quite noisy, and tells us little about d_i. The shrinkage coefficient c_i balances these two factors.

For instance, Rubin [12] models each $\hat{d}_i^{np}$ as normal, with mean d_i (since it is unbiased) and standard error σ_i:

$$\hat{d}_i^{np} \sim \mathcal{N}(d_i, \sigma_i) \quad (2)$$

This would be approximately the case if estimators $\hat{d}^{np}$ were difference-in-means or regression estimators from sufficiently large experiments. Then, he models the effects themselves as drawn from a normal distribution:

$$d_i \sim \mathcal{N}(\mu, \tau). \quad (3)$$

Under model (2)–(3),

$$c_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}. \quad (4)$$

When τ is large (so the true effects are very different from each other) and σ_i is small (so $\hat{d}_i^{np}$ is very precise), c_i is close to one—the partial pooling estimator $\hat{d}_i^{pp} \approx \hat{d}_i^{np}$—data are barely pooled across experiments at all. Conversely, when σ_i is large (so $\hat{d}_i^{np}$ is noisy) and τ is small (so the true effects are similar to each other), then c_i is close to zero, and $\hat{d}^{pp} \approx \mu$, the overall mean effect size, completely pooling data across experiments. In general c_i is in between zero and one, and the estimator $\hat{d}^{pp}$ partially pools information between the individual effect estimate $\hat{d}_i^{np}$ and the overall mean μ. The mean of the true effects μ and their variance τ are, of course, unknown, but they may be estimated from the data.

Unlike $\hat{d}_i^{np}$, $\hat{d}_i^{pp}$ is biased—it is shrunk towards the overall mean μ. To compensate for the bias, $\hat{d}_i^{pp}$ is less noisy than $\hat{d}_i^{np}$; its standard error is $\sqrt{c_i}\sigma_i$. Since $c_i < 1$, this is always less than $\hat{d}_i^{np}$'s standard error σ_i. Overall, [15] shows the root mean squared error (RMSE) of the estimates $\hat{d}^{pp}$, considered as a group, will be less than the RMSE of the individual unbiased estimates $\hat{d}^{np}$. This result is that it applies even when the causal estimates do not need to be related in any way.

When analyzing a set of A/B tests run inside intelligent tutors, estimates of the effects based on partial pooling will be more accurate, on average, than estimates that consider each test individually.

3. ANALYZING 22 EXPERIMENTS

How does partial pooling work in practice, in an authentic EDM setting?

The ASSISTments TestBed [7] allows education researchers to propose and conduct minimally-invasive A/B tests within the ASSISTments intelligent tutor. The TestBed infrastructure automatically publishes anonymized data from these experiments. Conveniently, [13] combined 22 of these datasets into one publicly available file. All 22 experiments were skill builders, which are problem sets designed to teach, or bolster, a specific topic or skill. Inside a skill builder, students are required to solve problems associated to that skill until mastery is achieved, typically defined as answering three questions in a row.

The dataset includes a number of student features and two dependent measures. In this paper, We will focus only on one dependent measure `complete`, a binary variable indicating completion of the skill builder, taking value 1 if the student achieved mastery or 0 if the student either stopped working before achieving mastery or exhausted all of the skill builder's problems without achieving mastery.

To estimate treatment effects conventionally, without pooling across experiments, we fit a separate logistic regression to each of the 22 experiments, regressing `complete` on an indicator for treatment condition.

$$Pr(complete = 1) = invLogit\left(\alpha_{expr} + \beta_{expr} Z\right) \quad (5)$$

Where $invLogit(\cdot)$ is the inverse logit function. The intercept α_{expr} and treatment effect β_{expr} (the log odds ratio of completion for the treatment vs the control condition) were estimated separately in each experiment $expr$.

To estimate effects using partial pooling, we re-fit (5) within a Bayesian multilevel logistic regression using the `rstanarm` package [14] in `R` [11]. That is, we assigned models $\alpha_{expr} \sim \mathcal{N}(\alpha_0, \sigma_\alpha)$ and $\beta_{expr} \sim \mathcal{N}(\beta_0, \tau)$, where hyperparamters α_0, β_0, σ_α and σ_β were estimated from the data using weakly-informative priors.

Figure 1 plots estimated treatment effects and approximate 95% confidence intervals ($\pm 2SE$) for the 22 experiments, using both the conventional no-pooling estimator and the partially-pooling estimator. The partial pooling shrunk the estimates quite a bit: while the no-pooling estimates ranged from approximately -1.3 to 0.6, the partial pooling estimates were all close to zero, ranging from -0.2 to 0.1. The estimated standard errors were also much smaller for the partially pooled estimators. The average standard error for the no-pooling estimates was 0.39, whereas the average standard error for the partial-pooling estimates was less than half that, 0.17. Finally, though two of the no-pooling estimates were statistically significant, with confidence intervals excluding zero, none of the partial-pooling estimates was.

Figure 2 plots the estimated standard errors from the two

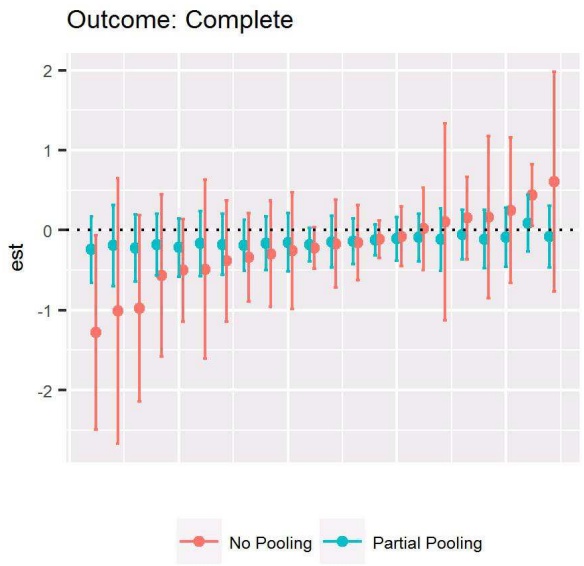

Figure 1: Partial-pooling an no-pooling treatment estimates and approximate 95% confidence intervals for the 22 experiments, arranged horizontally by the no-pooling treatment effect. The outcome was `complete`, and the treatment effects are log-odds ratios.

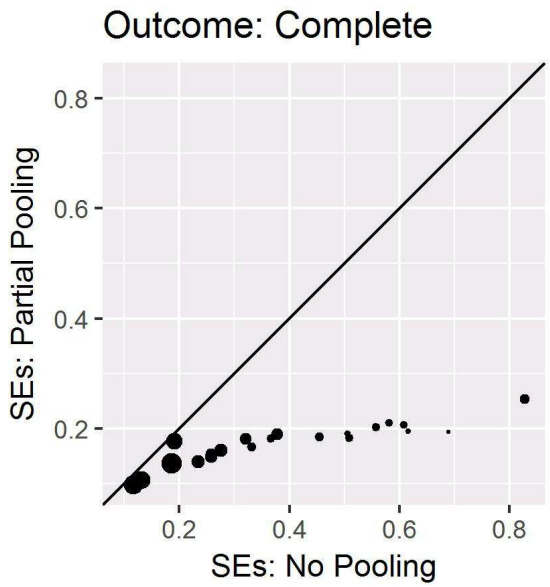

Figure 2: Partial-pooling vs no-pooling standard errors, with point size proportional to sample size in the experiment.

sets of estimates. The sizes of the points in the plot are proportional to experimental sample sizes. The partial-pooling standard errors are all smaller than those from the no-pooling estimates. However, the differences are not uniform. Experiments with large sample sizes and low no-pooling standard errors had partial-pooling standard errors that were only slightly smaller. As the sample sizes shrunk, both sets of standard errors grew. However, the no-pooling standard errors grew much faster. The largest difference in standard errors between the two methods was for studies with the smallest samples and the largest no-pooling standard errors.

4. A SIMULATION STUDY

Partial pooling worked as advertised when applied to the ASSISTments dataset, shrinking estimates towards zero and reducing standard errors, sometimes drastically—but did it get the right answers?

We ran a simulation study to investigate the performance of the partial-pooling estimator when the right answer is known.

4.1 Data Generating and Analysis Models

We simulated batches of $K = 20$ experiments each. Within a batch, sample sizes n varied from 20 to 115. Treatment Z was randomized to half of the subjects in each experiment. For each batch, outcomes Y were generated as

$$Y_i \sim \mathcal{N}(\alpha_{expr[i]} + \beta_{expr[i]} Z_i, \sigma_Y) \quad (6)$$

with random intercepts $\alpha_{expr} \sim \mathcal{N}(0, 1)$ and treatment effects $\beta_{expr} \sim \mathcal{N}(0, \tau)$, both varying at the experiment level. The between-experiment standard deviation of treatment effects τ varied between runs. It took the values of $\tau = 0$, corresponding to $\beta_{expr} \equiv 0$ across all experiments, and $\tau = \{0.1, 0.2, 0.5, 1.0\}$. When τ was positive but low, there was a treatment effect in every experiment, but nearly all effects were very small. Larger values of τ corresponded to more variance in the treatment effects, including some that were substantial. For every study, $\sigma_Y = 1$.

The 20 experiments in each batch were analyzed both separately, with no-pooling estimators, and jointly, with a partial-pooling estimator. Both estimators fit model 6 to each dataset to estimate treatment effects β_{expr}; however, the partial-pooling estimator additionally modeled $\beta_{expr} \sim \mathcal{N}(\beta_0, \tau)$ and $\alpha_{expr} \sim \mathcal{N}(\alpha_0, \sigma_\alpha)$.

For each value of τ we ran 500 iterations of 20 experiments each, producing 10,000 experimental datasets.

4.2 Simulation Results

Table 1 gives the results of the the simulation. The estimated standard errors and root mean squared errors of partial pooling estimates were consistently substantially lower than those of no-pooling estimates—partial pooling increased both accuracy and precision. The differences between the estimators diminished as the variance of true treatment effects, τ increased. This is predicted by (4): as τ increases relative to no-pooling standard errors σ, the shrinkage coefficient tends towards 1 and the the partial pooling estimate tends towards the no-pooling estimate. Intuitively, when τ increases various experiments become less informative about each other, so partial pooling decreases in value.

		τ				
	Pooling	0	0.1	0.2	0.5	1
SE	Partial	0.13	0.14	0.17	0.23	0.25
	None	0.26	0.27	0.26	0.27	0.26
|Bias|	Partial	0.00	-0.05	-0.08	-0.08	-0.05
	None	0.00	-0.00	0.00	0.00	0.00
RMSE	Partial	0.09	0.12	0.17	0.24	0.26
	None	0.28	0.27	0.27	0.28	0.27
Coverage	Partial	1.00	0.98	0.95	0.95	0.95
	None	0.95	0.95	0.95	0.95	0.95

Table 1: Average standard error (SE), bias magnitude, root mean squared error (RMSE), and empirical coverage of 95% confidence intervals (Coverage) for partial pooling and no pooling estimates for different values of τ.

Table 1 also shows that while the no-pooling estimates are unbiased, the partial pooling estimates are slightly biased towards zero, as expected, with the bias decreasing as τ increases. This bias does not cause undercoverage of 95% confidence intervals. Remarkably, for low τ, the partial pooling confidence intervals *over*-covered—more than 95% of the realized confidence intervals included the true parameter. The width of the confidence interval is four times the standard error, by construction—so partial-pooling confidence intervals were both substantially smaller and more often correct.

5. DISCUSSION

Partial pooling is a surprising, and surprisingly effective, technique to improve education sciences in the big data era. As educational technology allows A/B testing to proliferate, partial pooling is a method to use some of the oldest results in statistics—such as regression to the mean—alongside new Bayesian technology to improve the precision and accuracy of experimental estimates. When experiments can be analyzed in a group, the result is smaller confidence intervals with the same or higher coverage.

Partial pooling is a model based technique, and it remains to be seen how it performs when the model is severely misspecified. A host of Bayesian model checking procedures, including some suggested in [12], may be brought to bear on this question. In any event, most effect estimates are approximately normally distributed, by the central limit theorem, so methods based on normal theory will apply.

All code and data for this paper may be found at `https://github.com/adamSales/EDMpartialPooling`.

6. REFERENCES

[1] B. Efron and C. Morris. Stein's estimation rule and its competitorsâĂŤan empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[2] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[3] M. Feng, N. T. Heffernan, and K. R. Koedinger. Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In *International conference on intelligent tutoring systems*, pages 31–40. Springer, 2006.

[4] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[5] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[6] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

[7] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[8] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[9] Z. A. Pardos, M. Feng, N. T. Heffernan, and C. Linquist-Heffernan. Analyzing fine-grained skill models using bayesian and mixed effects methods. *Educational Data Mining*, page 50, 2007.

[10] T. Patikorn, D. Selent, N. T. Heffernan, J. E. Beck, and J. Zou. Using a single model trained across multiple experiments to improve the detection of treatment effects. In *Proceedings of the 10th International Conference of Educational Data Mining*, 2017.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[12] D. B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981.

[13] D. Selent, T. Patikorn, and N. Heffernan. Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 181–184. ACM, 2016.

[14] Stan Development Team. rstanarm: Bayesian applied regression modeling via Stan., 2016. R package version 2.13.1.

[15] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206, Berkeley, Calif., 1956. University of California Press.

[16] S. M. Stigler. The 1988 neyman memorial lecture: a galtonian perspective on shrinkage estimators. *Statistical Science*, pages 147–155, 1990.