Bayesian Estimation of Program-Specific Impacts in the HPOG Program

Stas Kolenikov[1], David Judkins[1]
[1] Abt Associates, 6130 Executive Blvd, Rockville, MD 20852

Abstract
Local Health Profession Opportunity Grants (HPOG) programs, funded by the Office of Family Assistance, Administration for Children and Families (ACF) of the U.S. Department of Health and Human Services, provide education, training, and support services to help transition low-income adults into healthcare occupations. ACF's Office of Planning, Research, and Evaluation oversaw an evaluation to assess the success of these HPOG programs and to provide program-specific estimates of impact. Direct estimation of the local average treatment effect (LATE) consists of simply comparing the means of outcomes for the local treatment and control groups to estimate local impacts. Unfortunately, most programs serve too few students to support estimation of local impacts. To overcome those issues, we developed a complementary set of Bayesian estimates of local impacts based on mixed effect models with random effects defined at the program level, and random slopes for the treatment indicator. Before preparing Bayesian estimates of local program effects for the second round of grants (HPOG 2.0), we demonstrated the techniques on the previous round (HPOG 1.0) evaluation data. In addition to allowing methods revisions without fear of accusations of p-hacking, demonstrating the techniques on HPOG 1.0 first allowed us to use the posterior distributions for components of variance as priors for components of variance for HPOG 2.0. As expected from general methodological considerations, Bayesian estimates of impact exhibited less variability than direct estimates did. Bayesian credible intervals were shorter, often by a factor of about 2 to 4, than the confidence intervals at the same coverage level. At the same time, a frequentist empirical Bayes (EB) analysis of the same mixed models produced confidence intervals that were half as long as Bayesian intervals, still, which highlights the importance of properly accounting for the uncertainty in the variance component estimation that EB methods cannot fully incorporate.

Key Words: observational studies, impact heterogeneity, generalized linear mixed models, Bayesian estimation, Hamiltonian Monte Carlo.

1. Health Profession Opportunity Grants Evaluation Project

Following on a first round of Health Profession Opportunity Grants (HPOG) Program awards in 2010 ("HPOG 1.0"), the Office of Family Assistance (OFA) of the Administration for Children and Families (ACF), within the U.S. Department of Health and Human Services, awarded a second round of 32 five-year grants ("HPOG 2.0") in 2015. Local HPOG programs provide education, training, and support services (including financial and other assistance) to Temporary Assistance for Needy Families (TANF) recipients and other low-income adults for occupations in the healthcare field that pay well and are expected to either experience labor shortages or be in high demand. ACF's Office of Program Research and Evaluation (OPRE) awarded a contract to Abt Associates to evaluate the performance of the grantees. Of the 32 grants, five were awarded to American Indian tribes or affiliated tribal organizations. The other 27 grantees designed and implemented 38 distinct local programs that are the focus of this paper. Klerman, Judkins, and Locke (2019) laid out a plan for the evaluation of these 38 programs on a pooled basis. Subsequent to the publication of that plan, OPRE requested research on how to assess the impacts each of the 38 local programs separately. This paper shares results on that methodology. The results of both the pooled and local impact estimates are forthcoming (Klerman et al, forthcoming). Both the pooled and local evaluations aim to estimate the impact of access to the local HPOG 2.0 program rather than the impact of acceptance of the offer. This choice is mostly driven by the desire of the evaluators to remain within the design-based inference framework. Evaluating the impact of active engagement with an intervention invariably involves much stronger assumptions than are required to estimate the impact of the offer of access to the intervention. In addition to this motivation, a cogent argument can be made that this estimate is most relevant to policy analysts since no one can be forced to engage in a voluntary intervention like HPOG. For an intervention to be successful, it must make itself attractive to potential beneficiaries.

The impact analysis uses the data from a Short-term Follow-up Survey about 15 months after randomization and administrative data from the National Directory of New Hires and the National Student Clearinghouse through two and half years after randomization. The analyses in this report were pre-specified in the Analysis Plan for Short-term Impact Report (STIR) (Judkins, et al. 2020) and registered with the Open Science Framework and the Registry of Efficacy and Effectiveness Studies.

One of the Impact Evaluation's research questions is estimating the impact of each of the programs implemented by HPOG 2.0 grantees. The primary motivation for this activity is to be able to provide feedback to the grantees, and to inform the future cohorts of program participants and administrators. This research question thus has a prospective evaluation agenda. Unfortunately, most programs serve too few students to support rigorous local evaluations. Unbiased estimates of local effects can be produced with very small sample sizes, but the predictive value of these estimates for future cohorts may be low. Bayesian methods offer a possible approach to this challenge.

Before preparing Bayesian estimates of local program effects for HPOG 2.0, we first applied the techniques on the HPOG 1.0 data. In addition to allowing methods revisions without fear of accusations of p-hacking, demonstrating the techniques on HPOG 1.0 first allowed us to use the posterior distributions for components of variance as priors for components of variance for HPOG 2.0. While one of the selling points of Bayesian approach is incorporation of the existing information as prior distributions, our perception is that this is done surprisingly seldom in practice. Our project is a happy exception: the outcomes are defined and measured in exactly the same way between HPOG 1.0 and HPOG 2.0, and the information learned in HPOG 1.0 can be used directly in HPOG 2.0. Passing the historic information on impact heterogeneity allowed for sharper inferences about the local program effects of HPOG 2.0 programs.

2. Methodology

Our approach generally follows that of Meager (2019), who addressed a similar issue of understanding the heterogeneity in microcredit program effects. The supplementary estimates are obtained with assistance of a statistical model, namely a mixed logistic regression model with key HPOG 1.0 outcomes as dependent variables; treatment indicator and predictors as proposed in HPOG 2.0 analysis plan; and program-specific intercepts and treatment effects.

2.1 Choice of estimators

We used three different methods to estimate local treatment effects and associated error bands:

1. A frequentist maximum likelihood estimation of the model parameters followed up by empirical Bayes estimation of program-specific intercepts and treatment effects; and

2. A full Bayesian model which estimates everything at once, and simulates values of the model parameters and program-specific intercepts and treatment effects from their joint posterior distribution.

3. A frequentist design-based approach, called direct estimates, i.e., the program-specific differences in mean outcome between the treatment and the control groups that does not rely on any statistical models.

There are three aspects of Bayesian modeling that make Bayesian approach appealing for this application.

First, in comparison to the direct estimates, both the empirical Bayes method (#1) and the full Bayes method (#2) produce more accurate estimates of local program effects for prospective cohorts. They do this through a combination of “borrowing strength” across programs and “shrinking.” The jargon “borrowing strength” means assuming that structural relationships between baseline covariates and outcomes are universal. These relations do not have to be causal; the model does not require causality, and simply exploits associations between variables. Based on this assumption, the method can predict an average outcome for each program based on the profile of participants at baseline. Modeling that only uses person-level covariates only captures idiosyncratic variability due to local student demographics measured at baseline. There are many other possible sources of idiosyncratic variability such as program design, program implementation, and the skills of local staff. These aspects of implementation are hard to encode with quantitative variables. With large enough sample sizes, the direct design-based estimates (#3) fully capture idiosyncratic variability but are also very noisy because of the low sample sizes. By exploiting similarities in outcomes between similar units through the statistical model, both methods #1 and #2 shrink the direct estimates of method #3 to projections based on the assumption of universal structural relationships. The improvement in accuracy of methods #1 and #2 relative to method #3 depends on the strength of the structural relationships, local sample sizes, and the importance of idiosyncratic effects.

Second, given adequate computing power, estimation of error bands on local effects is much simpler with the (mostly computational) full Bayesian approach than with the (mostly analytical) empirical Bayesian approach. Although the theory has been worked out to construct fairly well-behaved error bands on local effects with the empirical Bayesian approach (Lahiri 2003), this theory has not been built into any of the standard statistical analysis systems. As a result, most practitioners who use the empirical Bayesian approach assume that the variance components are known without error. This runs the danger of underestimating variability of the estimates, which we will see in our results. In contrast, full Bayesian methods create a series of simulated values, or draws, from the statistical model that can be transformed to any statistic of interest. (Distributions of statistics produced that way are referred to as posterior distributions in Bayesian inference. The name indicates that these are distributions of statistics after we have observed the data; Bayesian inference also requires formulation of prior distributions reflecting our knowledge of the parameter values before we observe the data, e.g. from earlier studies.) Such draws are made for the model parameters, such as fixed effect coefficients and random effect variances; for the random effects that only participate in the model indirectly; and even for a hypothetical distribution of the treatment effect for a new, yet unobserved, program. Availability of such draws means two things. First, uncertainty in parameter estimates can be quantified as the variance across the draws. Second, complicated sample statistics such as small area estimates of treatment effects at the program level can be computed draw-by-draw, and their posterior distributions can thus be observed directly.

Until recently, the computing power to implement a full Bayesian approach was hard to acquire. However, recent software developments along with hardware improvement have made it feasible. We have used Hamiltonian Monte Carlo Markov chain methods (Neal 2011) implemented in statistical modeling package Stan with R interface RStan (Carpenter et al 2017). It provides extremely fast simulation of model parameters but requires a special set up of the computing platforms (access to the C compiler). While this was implementable with the survey data on the Abt computing platform, it could not be implemented on the ACF computing platform, where we used the previous generation of computational methods, Metropolis-Hastings algorithms implemented in Stata 16 (StataCorp 2019). Note that earlier versions of Stata did not support all of the necessary functionality.

Third, the use of Bayesian framework allows the researcher to utilize information on possible values of model parameters available prior to the study, to be used in estimation of the model parameters. This is done in the form of specifying the prior distributions for the model parameters. In modeling HPOG 1.0 data, we used priors that were only motivated by the plausible range of parameters (e.g., odds ratios rarely exceeding 1 for the binary outcomes, and changes in earnings rarely exceeding $1,000 per quarter.) In the subsequent work with HPOG 2.0 data, we used the summaries of the posterior distributions from the HPOG 1.0 Bayesian analysis as the priors for the HPOG 2.0 analysis for the critical parameters of treatment heterogeneity, as explained below.

2.2 Generalized linear mixed models: summary

Description of the statistical model presented here follows our initial proposal in the Analysis Plan for the HPOG 2.0 National Evaluation Short-Term Impact Report (Judkins, Klerman and Locke 2020). The predictors to be used were pre-selected and registered in advance. The principal model is a generalized linear mixed model:

$$\theta_{ij} = x_{ij}^T\beta + \gamma T_{ij}\, u_i + \tau_i T_{ij}$$

$$y_{ij} \sim \begin{cases} N\left(\theta_{ij}, \sigma_0^2\right): & y_{ij} \text{ is a continuous outcome} \\ \text{Bernoulli}\left(1/\{1+\exp[-\theta_{ij}]\}\right): & y_{ij} \text{ is a binary outcome} \end{cases}$$

$$\begin{pmatrix} u_i \\ \tau_i \end{pmatrix} \sim N\,(0, G)$$

Notation is as follows:

- θ_{ij} is a linear predictor of the outcome
- x_{ij} are (individual) control variables / predictors (including the intercept)
- β is the vector of regression coefficients for the control variables
- γ is the overall treatment effect
- T_{ij} is the person-level treatment indicator
- u_i is the program level effect unexplained by the control variables (i.e., by how much the program participants in the control arm fare better or worse than those in the control arm in other programs)
- τ_i is the program level treatment effect; the sum $u_i + \tau_i$ describes by how much the program participants in the treatment arm fare better or worse than those in the treatment arm other programs
- y_{ij} is the outcome; the distribution of this outcome is that of an exponential family

The random effects u_i (random intercept) and τ_i (random treatment effect slope) are assumed to follow a joint bivariate distribution with mean zero and variance-covariance matrix G, so that g_{11} is the variance of the random intercepts, g_{22} is the variance of random slopes, and g_{12} is the covariance between the two. With arbitrary values of g_{11}, g_{12} and g_{22}, there is a danger that the resulting G matrix composed of these values is not positive definite. Different packages adopt different parameterization strategies, including Cholesky decomposition that leads to uninterpretable coefficients; transformations of parameters such as taking logs of variances and hyperbolic arctan transformations of correlations; or decomposition of the matrix as the product of a diagonal matrix of variances vis-à-vis a proper correlation matrix.

If the outcome is continuous, the model is the linear mixed model with random slopes for treatment (the outcome is normal with conditional mean $x_{ij}^T\beta + u_i + \tau_i T_{ij} \equiv \theta_{ij}$ and residual variance σ_0^2; self-reported income was log-transformed to bring it closer to normality); and if the outcome is binary, the model is the logistic mixed model with log odds ratio given by $x_{ij}^T\beta + u_i + \tau_i T_{ij} \equiv \theta_{ij}$. While there is no theoretical reason to disfavor the use of either person-level or program-level predictors as x_{ij}, we only used person-level covariates in this application. The most important parameters, in the evaluation context, are the overall treatment effect γ and its heterogeneity g_{22}. If the latter heterogeneity is zero, the treatment effect estimate $\hat{\gamma}$ should be expected to have tremendous external validity, in the sense that all current programs produced nearly identical estimates, and it is likely that future sites implementing this program can reasonably be expected to exhibit performance very similar to the common estimate $\hat{\gamma}$. With nonzero heterogeneity of treatment effects g_{22}, one needs to adjust expectations on the program performance to account for that variability. We implicitly assume that the set of local programs developed by both HPOG 1.0 and HPOG 2.0 grantees are both random samples of the universe of programs that might be implemented in the future. While this is obviously a strong assumption, the consequence of making this assumption is wider, i.e. more conservative, confidence intervals for gamma, something that is likely to be beneficial for the drawing of public policy conclusions from the report. The key reference for this methodology and application of the model is Meager (2019).

2.3 Computation

Maximum likelihood estimation of this model proceeds by writing out the full likelihood of the sample, where the likelihood contributions of individual observations involve numeric integration of the random effects (e.g. by Gaussian quadrature) or approximating the curvature at the mode of this distribution (PQL method). It is available in standard statistical software, such as SAS `PROC MIXED/PROC GLIMMIX`; Stata official `mixed` and `meglm` commands as well as the user contributed `gllamm` package (Rabe-Hesketh et al 2005); R `library(lme4)` (Bates et al 2015) and `library(nlme)` (Pinheiro et al 2020).

Bayesian estimation proceeds by formulating prior distributions of model parameters, and updating them with the available data using Bayes theorem. In the latter step, Markov chain Monte Carlo (MCMC) samples are taken from the posterior distribution that combines prior distributions and the data. Among a number of available packages, the most modern one that achieves the highest speed is Stan (Carpenter et al 2017). Just as the previous generations of MCMC algorithms used ideas from computational physics, such as Gibbs sampling that has roots in statistical physics, Stan utilizes ideas of Hamiltonian dynamics (Neal 2011) to create highly effective draws from posterior distributions. A very high computational speed is achieved by first converting the model code into interim C code, and then compiling that code into high performance executable binaries. With the Short-term Follow-up Survey data, we used the R interface to Stan provided by RStan package (Stan Development Team 2020). With NDNH data, setting up RStan proved impossible due to security restrictions (namely not being able to invoke the C code compilers), so estimation was performed in Stata using the previous conceptual generation of MCMC samplers (Gibbs samplers and Metropolis-Hastings algorithm).

The prior distributions can be based on earlier estimates if those are available. This is what was done for HPOG 2.0, namely (approximations to) the HPOG 1.0 posteriors were used to formulate the priors (and thus get sharper results). For HPOG 1.0, such information was not available, and we instead used priors based on the known ranges of values. The quarterly earnings are expected to be in the range of thousands of dollars for the HPOG population, with rare exceedances into low tens of thousands, so the coefficients are expected to be in the range of hundreds and thousands. The log odd ratios for the binary outcomes can be expected to be in the range of at most –5 to 5, so the coefficients and variance components are expected to be on the scale of about 1.

Since estimation was conducted on two different platforms, we had to use different prior formulations for the self-reported survey outcomes and for the NDNH earnings.

In the binary outcome models for the survey self-reported data estimated using RStan, prior distributions for the regression coefficients were Laplace (double exponential distribution) with the center parameter at zero, and scale parameter of 0.25 for binary outcomes and log-transformed income outcome. The use of Laplace priors leads to Bayesian analogue of the least absolute shrinkage and selection operator, lasso, a popular model selection and regularization tool, and reflect our expectation that most regression parameters are close to zero. The scale parameters were chosen to correspond to the expected ranges of the coefficient estimates.

Prior distributions of variance components were half-Cauchy with scale of 1. Such priors allow variance components to take both values close to zero, and large values, reflecting our lack of specific expectations about those variances. The scale of that distribution, however, is chosen to be commensurate with the overall scale of predicted values of the linear index θ_{ij}. As our posterior plots show, the data strongly overwhelm this prior. The prior for correlation between the random intercept and random treatment slope was the Lewandowski-Kurowikca-Joe (LKJ) prior with shape parameter of 1, which is equivalent to uniform [–1,1]. The full covariance matrix G is obtained by multiplying through the vector of variance components and the correlation matrix.

For the NDNH earnings that were available on a different platform, we had to rely on the implementation of Metropolis-Hastings algorithms in Stata, which is the previous generation of MCMC algorithms. In that environment, the choice of the prior distributions implemented in the software was more limited. We used non-informative multivariate normal distributions for the priors for regression coefficients, and inverse Wishart distribution with moderate degrees of freedom, namely 6, and the identity scale matrix as the prior for *G*.

3. Results

We communicate the results mostly with graphs. We first discuss how to interpret these graphs with the aid of a generic outcome. We present three types of graphs:

1. Caterpillar plots of the program-specific treatment effects. When examining one of these plots, the reader is able to determine the estimate of the effect of a specific program, place it in the context of other programs, understand the precision of the estimate, and understand the sensitivity of the estimate to methodological decisions.

2. Overlaid prior and posterior distribution of the standard deviation of the treatment effects random slopes. In program evaluation context, this is the most important parameter in the entire Bayesian modeling effort. By comparing the prior and posterior distributions, the reader can get some sense of the importance of the prior.

3. Anticipated efficacy of the program at a new site. This posterior predictive distribution directly demonstrates the magnitude of the effect and the likelihood that a new program funded by the same stream with the same regulations under the same general economic conditions is beneficial to its participants in terms of a given outcome.

3.1 Caterpillar plots

While the latter two kinds of plots simply visualize (simulated, approximate) probability distributions of variables of interest, caterpillar plots are somewhat idiosyncratic to small area estimation models (with some similarities found in forest plots in the meta-analysis literature). Consider the following plot in Figure 1. These are direct survey estimates (differences of means by arm) of treatment effect in educational attainment. The estimates are sorted by the point estimate (the dot in the middle of the confidence interval). Three estimates are negative although not significant; the largest ones are about 0.3 (i.e. 30 percentage points). Program names are masked because the grantees to HPOG 1.0 were promised that identifiable statistics about their performance would not be published. The 95% confidence intervals are also depicted. The half-width of confidence intervals vary from about 0.1 to about 0.4. About half of the confidence intervals contain zero.

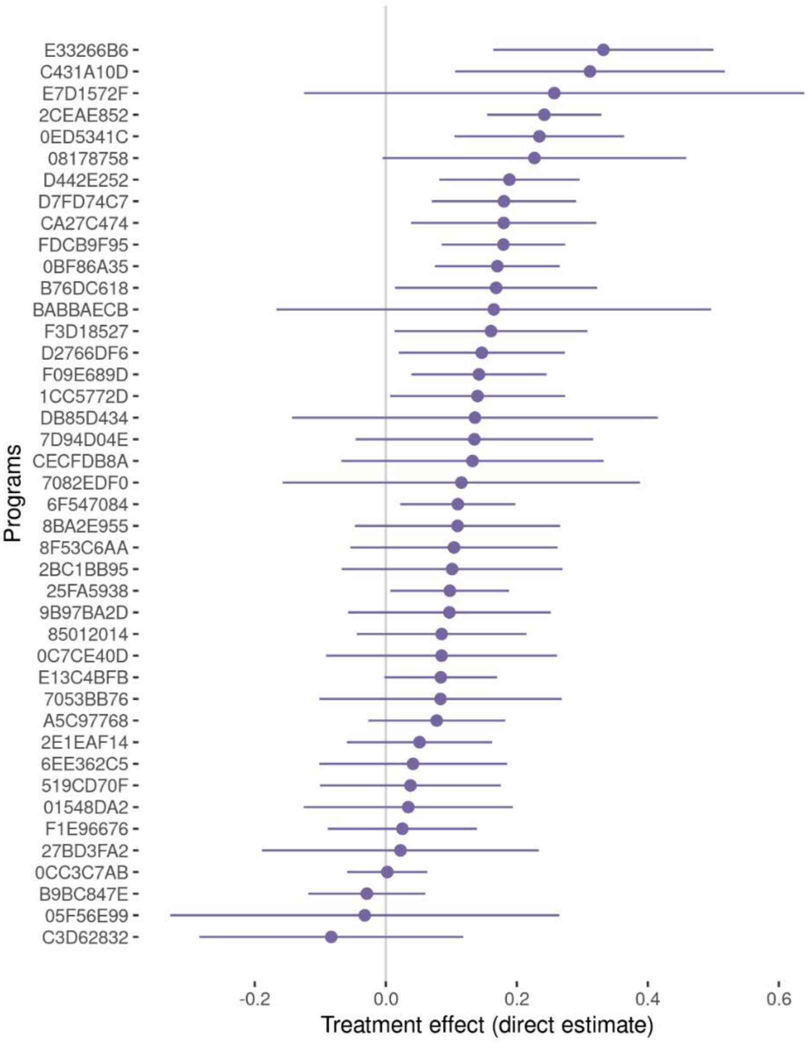


Figure 1. Example of a caterpillar plot: direct estimates.

Source: Abt Associates modeling of HPOG 1.0 data.

Let us now present a caterpillar plot for the same outcome but using Bayesian small area estimation methods (Figure 2). It uses the same principles: programs are sorted by estimates, and credible intervals are shown. Note the change of scale of the x-axis: the range in Figure 1 is from –0.3 to 0.6, while the range in Figure 2 is from about –0.1 to about 0.25, reflecting shrinkage of Bayesian estimates. All estimates are positive, although the 8 out of 10 smallest ones are not significantly different from zero. The estimates range from 0.03 to 0.18, exhibiting much smaller spread. The half-length of credible intervals is 0.05 to 0.10, much shorter than that of the direct estimates in Figure 1.

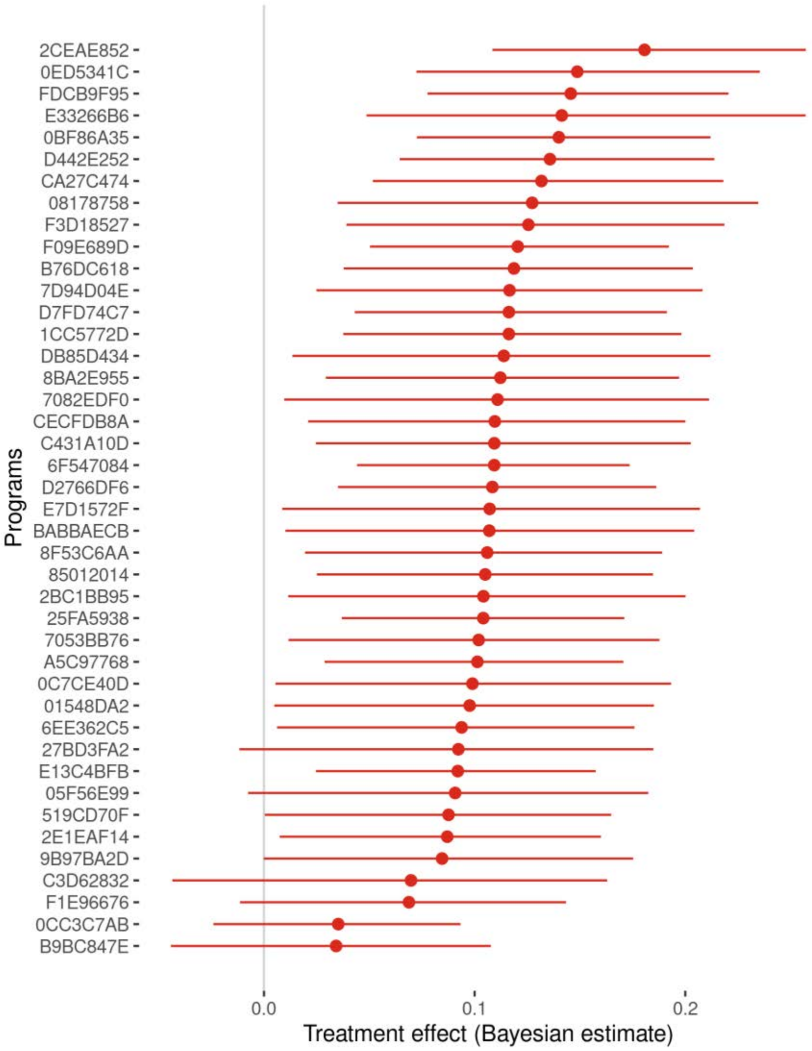


Figure 2. Example of a caterpillar plot: Bayesian small area estimates.

Source: Abt Associates modeling of HPOG 1.0 data.

Finally, we also considered another frequentist small area estimation method that is based on maximum likelihood estimation of random effect models and empirical Bayes prediction of random effects. The results are presented in Figure 3.

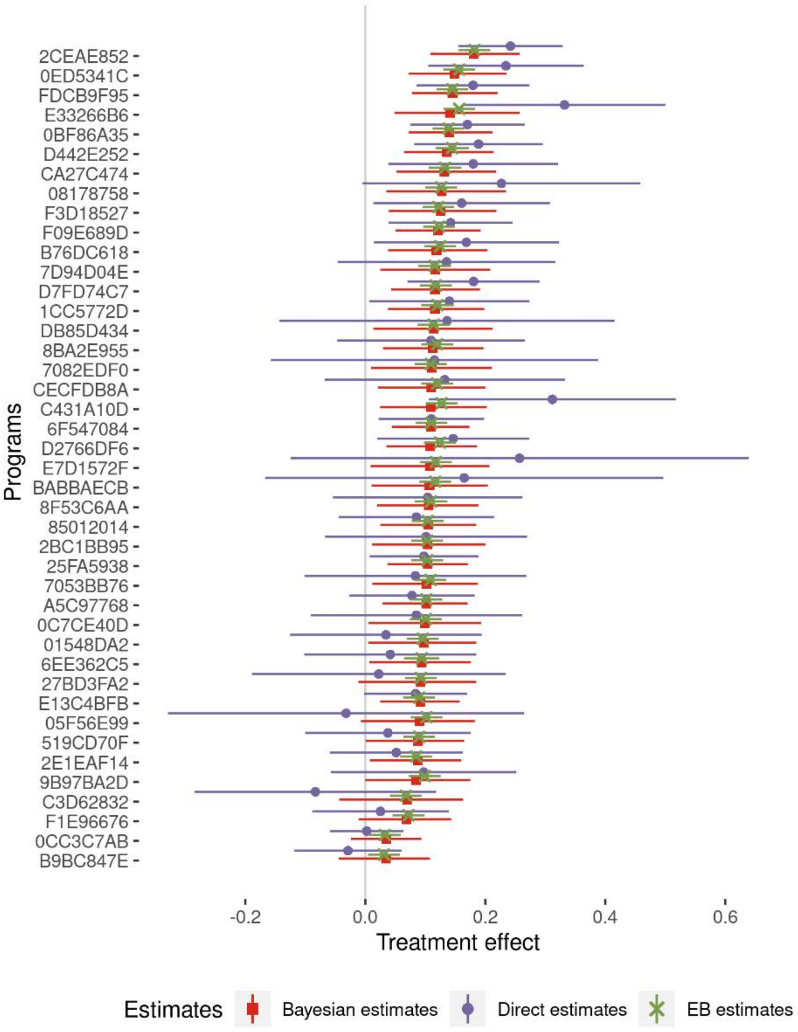


Figure 3. Combined direct, mixed model assisted empirical Bayes, and Bayesian caterpillar plots.

Source: Abt Associates modeling of HPOG 1.0 data.

Blue-purple lines with circles are direct estimates, i.e., the differences in mean outcome within a program. Red lines and squares are Bayesian estimates. Green lines with an X mark are small area estimates based on the mixed model estimated by maximum likelihood, and empirical Bayes predictions of random effects. The range is the 95% confidence interval, and the marked point is the estimate. Note that the green X are almost on top of the red squares. This means that the Bayesian and frequentist mixed modeling approaches yield nearly identical estimates of local program impacts. However, the error bands are dramatically different. As mentioned above, the EB estimate as currently implemented is known to not account for the sampling errors in the estimate of the random effect variances; as a result, the confidence intervals are too short. The agreement of the model-based and the direct estimates is in that Bayesian credible intervals and the EB confidence intervals are generally contained within the confidence intervals of the direct estimates and have nonzero overlap for all areas.

Given the deficiencies of the EB approach, we treat this method as a sensitivity check and a comparison to other methods, especially given that this method is actually used in the small area estimation literature; but we do not recommend using it for policy. Rao and Molina (2015) discuss the mean squared error (MSE) and its estimation for linear models in sections 7.1 and 7.2. They introduce useful quantities of the contributions to the MSE: the term g_{1i} due to prediction of the area (program) effect, the term g_{2i} due to estimating the regression coefficients, and the term g_{3i} due to estimating the variance components. The first two terms can be obtained by moderately complicated algebra, and they are encoded in the confidence intervals presented on the caterpillar plots. Estimation of the third term is complicated. Rao and Molina (2015) cite methods that rely on specific distributional assumptions (normality of the random effects), method of moments estimators of the variance components, and parametric bootstrap methods for MSE estimation (Lahiri 2003). It appears that the best performing methods are the double bootstrap methods, where (1) a bootstrap sample is constructed first, and (2) another bootstrap loop is used to estimate the requisite variances within that sample. These methods impose an exponentially higher computational burden, where the underlying mixed logistic regression models would need to be run hundreds of thousands of times. We decided against this approach, especially given that the Bayesian approach was sufficiently simple to implement and provided additional information helpful in the program evaluation context, such as the anticipated efficacy forecast, that the frequentist method would not be able to.

3.2 Passing distributions from HPOG 1.0 to HPOG 2.0

The distribution of $\sqrt{g_{22}}$, the standard deviation of the random slope of treatment for the first round is presented in Figure 4 for the same outcome as the caterpillar plots in Figures 1–3.

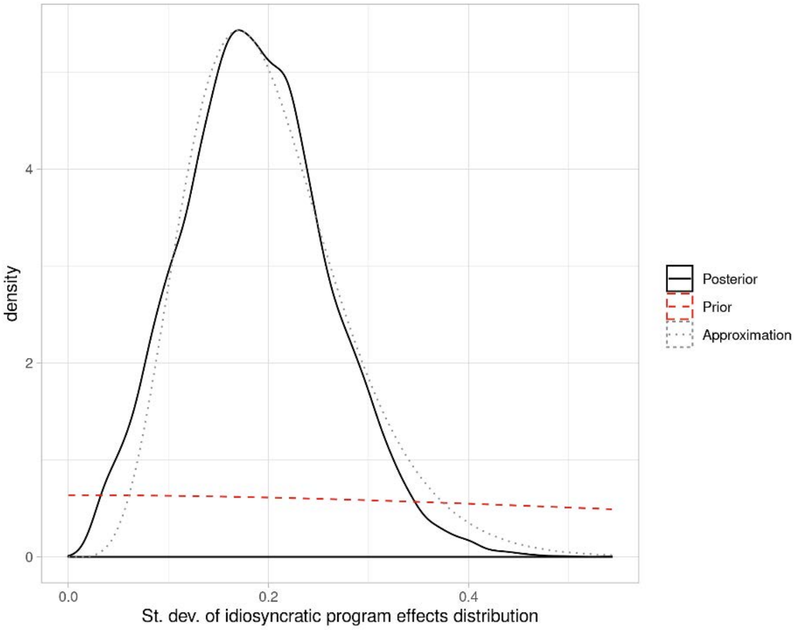


Figure 4. Prior vs. posterior distribution of $\sqrt{g_{22}}$, and a gamma approximation, for HPOG 1.0 data and analysis. Prior distribution: half-Cauchy with scale of 1; posterior approximation: gamma with shape α=6.55, rate β= 0.0307.

]

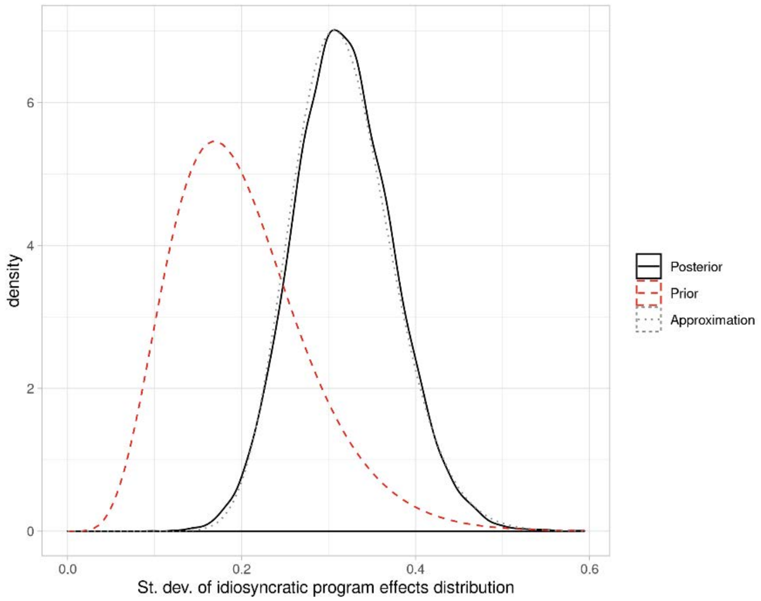
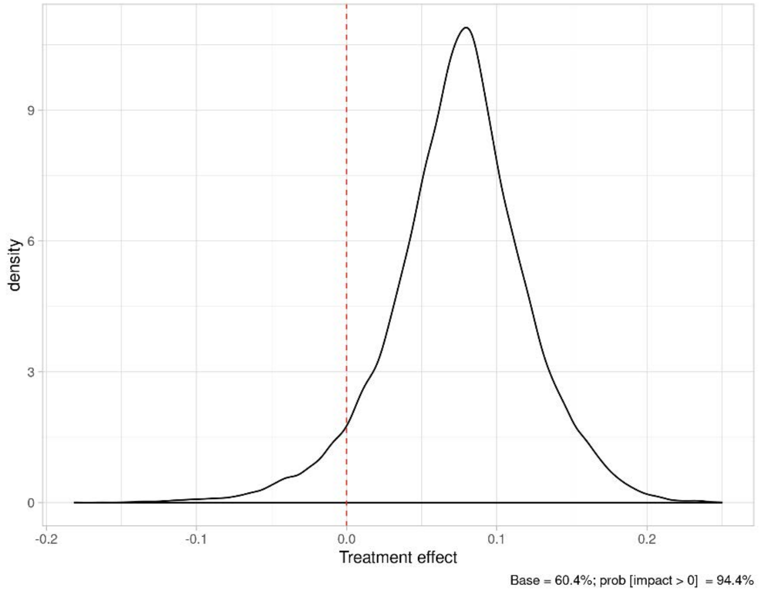


Figure 5. Prior vs. posterior distribution of $\sqrt{g_{22}}$, and a gamma approximation, for HPOG 2.0 data and analysis. Prior distribution: gamma with shape α=6.55, rate β= 0.0307; posterior distribution approximation: gamma with shape α= 30.25, rate β= 0.01048.

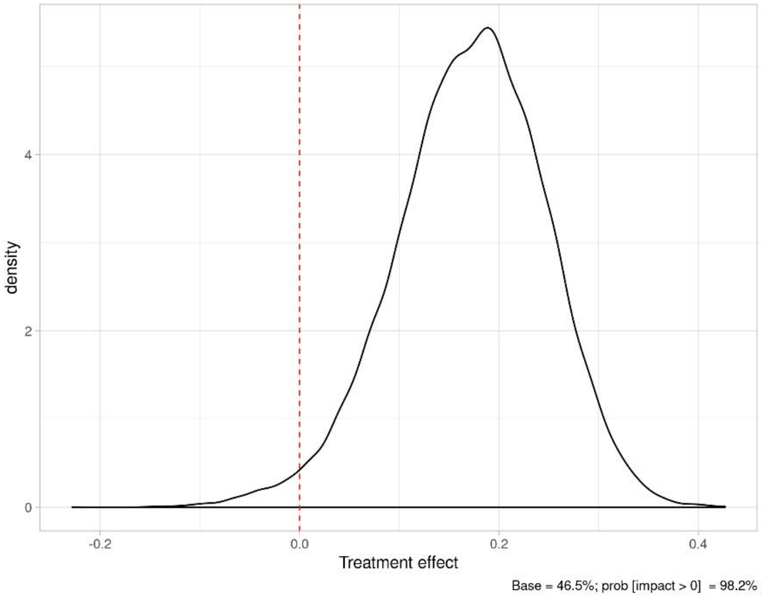
The results for the other two parameters of the random effect covariance matrix G, the random intercept variance g_{11} and the correlation parameter $g_{12}/\sqrt{g_{11}g_{22}}$ have less prominence in the substantive evaluation terms than the variance of the treatment effect slopes that has an important interpretation of the extent of external validity and reproducibility, but these parameters are also important in terms of providing priors for the subsequent HPOG 2.0 analysis. The suggested parameters for the random intercept variance are the method of moments estimates for gamma distribution (which is a generalization of the χ^2 distribution that variance estimates follow for the case of the normal data, allowing for arbitrary scale and shape/degrees of freedom.) The degrees of freedom parameter for g_{11} ranged from about 20 to about 50. The moment estimate for the LKJ posterior shape parameter η can found from the relation to the variance, $\text{Var}[r] = \Gamma(\eta+\frac{1}{2})/2\Gamma(\eta+\frac{3}{2})$. However, since the distribution is symmetric around zero, it is does provide a good approximation to the posterior distributions encountered in practice. For the purposes of passing the information about correlation concentration to the future rounds, the LKJ shape parameter was instead estimated from the posterior distribution percentile to match the 90th percentile of absolute values. As the bivariate correlation matrix has determinant $1-\rho^2$, and the LKJ distribution is characterized by $f(\boldsymbol{C},\eta) \propto |\boldsymbol{C}|^{\eta-1} = (1-\rho^2)^{\eta-1}$ where $\boldsymbol{C}$ is the resulting correlation matrix, the distribution of ρ^2 is $\text{Beta}(1,\eta-1)$ so the required parameter estimates can be obtained from an inverse incomplete Beta function. For the outcomes studied, the HPOG 1.0 posteriors corresponded to the shape parameter ranging from 3.0 to 6.9.

3.3 Predictions at an unobserved site

Bayesian estimation allows for easy MCMC simulation of the predicted model performance at a new site. For the k-th iteration of the chain, the prediction can be formed as of $\gamma^{(k)} + \tau_{\text{new}}^{(k)}$ where the hypothetical program effect at a new, yet unobserved location $\tau_{\text{new}}^{(k)}$ is drawn from $N(0, g_{22}^{(k)})$. The resulting distributions, being mixtures of normals, typically have tails heavier than normal. A substantively important predictive quantity is the posterior prediction of the program having a positive impact; for wave 1, it was estimated at 94.4%, and for wave 2, at 98.2%. As the variance components distribution in wave 2 is characterized by a much higher number of degrees of freedom, the wave 2 distribution is closer to normal.



(i) HPOG 1.0



(ii) HPOG 2.0

Figure 6. Predicted impacts at a new site: (i) HPOG 1.0; (ii) HPOG 2.0.

Acknowledgements

The authors would like to thank Jacob Klerman and Gretchen Locke of Abt Associates and Nicole Constance of OPRE for helpful comments that helped sharpen the AAPOR and the JSM presentations, and the proceedings paper.

References

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1). 10.18637/jss.v076.i01

Judkins, D. R., Klerman, J. A., and Locke, G. (2020). Analysis Plan for the HPOG 2.0 National Evaluation Short-Term Impact Report. *OPRE Report # 2020-07*. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Judkins, D. R., Klerman, J. A., and Locke, G. (2020). *Analysis Plan for the HPOG 2.0 National Evaluation Short-Term Impact Repo*rt. OPRE Report 2020-07. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/report/national-and-tribal-evaluation-2nd-generation-health-profession-opportunity-grants-0.

Klerman, J. A., Judkins, D. R., and Locke, G. (2019). *Impact Evaluation Design Plan for the HPOG 2.0 National Evaluation.* OPRE Report 2019-82. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/report/national-and-tribal-evaluation-2nd-generation-health-profession-opportunity-grants-1.

Klerman, J. A., Judkins, D. R., Prenovitz, S., and Locke, G. (Forthcoming). *Health Profession Opportunity Grants (HPOG 2.0) Short-Term Impact Report.* Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Lahiri, P. (2003). On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation. *Statistical Science*, 18 (2), 199–210.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100 (9), 1989–2001.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Association: Applied Economics*, 11, 57-91.

Neal, R. (2011). MCMC Using Hamiltonian Dynamics. Chapter 5 in: *Handbook of Markov Chain Monte Carlo*, edited by Steve Brooks, Andrew Gelman, Galin L. Jones and Xiao-Li Meng. Chapman & Hall/CRC.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team (2020). nlme: Linear and Nonlinear Mixed Effects Models. *R package*, version 3.1-145, available from https://CRAN.R-project.org/package=nlme.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128 (2), 301-323.

Rao, J. N. K., and Molina, I. (2015). *Small Area Estimation*. 2nd edition. Hoboken, NJ: Wiley.

Stan Development Team (2020). RStan: the R interface to Stan. *R package*, version 2.19.3. http://mc-stan.org/.

StataCorp (2019). *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.