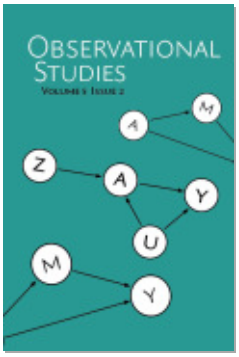
Examining treatment effect heterogeneity using BART

Nicole Carnegie, Vincent Dorie, Jennifer L. Hill



Observational Studies, Volume 5, Issue 2, 2019, pp. 52-70 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2019.0002*

Observational Studies 5 (2019) 52-70 Submitted 9/18; Published 8/19

Examining treatment effect heterogeneity using BART

Nicole Carnegie **nicole.carnegie@montana.edu**
Montana State University, Bozeman, MT, USA
Vincent Dorie **vjd2106@columbia.edu**
Columbia University, New York, NY, USA
Jennifer L. Hill **jennifer.hill@nyu.edu**
New York University, New York, NY, USA

Keywords: Causal Inference, Bayesian Additive Regression Trees, Treatment Effect Modification, Group-structured Data

1. Methodology and Motivation

We were presented with the challenge of estimating causal effects using simulated data that was intended to roughly mirror "preliminary data extracted from the National Study of Learning Mindsets." In particular, we were asked to address three research goals:

1. *Was the mindset intervention effective in improving student achievement?.*
2. *Researchers hypothesize that the effect of the intervention is moderated by school level achievement ($X2$) and pre-existing mindset norms ($X1$). In particular there are two competing hypotheses about how $X2$ moderates the effect of the intervention: Either it is largest in middle-achieving schools (a "Goldilocks effect") or is decreasing in school-level achievement.*
3. *Researchers also collected other covariates and are interested in exploring their possible role in moderating treatment effects.*

We discuss our approach to these three research goals as well as a summary of our results.

1.1 Assumptions

Given that the simulated dataset was based on data from a large-scale randomized experiment, the Learning Mindsets study, we were hopeful that the simulated data satisfied ignorability for the research questions posed. To be conservative, we assumed that ignorability was conditional on the full set of observed covariates—that is, $Y(0), Y(1) \perp Z \mid X$ (Rubin 1979). In the post-workshop analyses we examined the sensitivity of our estimates to violations of ignorability and were satisfied that it was not an unreasonable assumption.

Given this ignorability assumption, any analysis would require appropriate conditioning on covariates to achieve unbiased estimates of $\mathrm{E}[Y(0) \mid X]$ and $\mathrm{E}[Y(1) \mid X]$. We had two strategies for avoiding strong parametric assumptions. First, we checked that each variable

we controlled for satisfied balance and overlap. This helped ensure that empirical counterfactuals existed for all observations. Second, we used a very flexible modeling strategy for estimating these conditional expectations.

At different stages in our analysis, we made several different types of modeling choices with respect to the grouped data structure. Each has its own set of assumptions. When we represented this structure through school-specific fixed effects, α, our ignorability assumption generalized to $Y(0), Y(1) \perp Z \mid X, \alpha$. When we instead modeled school-level variation as varying intercepts—or "random effects"—we imposed the additional assumption that the random effects were uncorrelated with the (school-level aggregates of) covariates and treatment indicator. This would be violated if an unobserved school-level covariate was predictive of both school-level treatment rates and mean response.

We had no way of knowing whether SUTVA was satisfied. We performed analyses under the assumption that it was satisfied.

1.2 Choice of BART as the foundation of our approach

Without information about the true parametric form of the response surface, we opted for a method that flexibly fit the response surface. Recent evidence demonstrates the advantages of machine learning algorithms as an approach to causal effect estimation (for instance, Hill 2011; Dorie et al. 2018). Within this class of estimators, we prefer automated algorithms that have been integrated into Bayesian inferential frameworks. This combination allows for uncertainty quantification and is more flexible in accommodating several other complications such as grouped data structures and missing outcome data. One such modeling strategy, based on Bayesian Additive Regression Trees (BART; Chipman et al. 2007, 2010), already has a proven track record of superior performance in causal inference settings (Hill 2011; Hill et al. 2011; Hill and Su 2013; Dorie et al. 2016; Kern et al. 2016; Wendling et al. 2018). Here we briefly introduce BART and its uses for causal inference.

Bayesian Additive Regression Trees. The BART algorithm consists of a sum-of-trees model and a regularization prior. The prior avoids overfit by specifying the number of trees, the probability distribution for the size of each tree, the shrinkage applied to the fit from each tree, and the degrees of freedom for the prior distribution for the residual standard error. Interested readers can find more information on the model, prior, and fitting algorithms in Chipman et al. (2007, 2010). The key point is that BART can be used to flexibly fit even highly nonlinear response surfaces, which is consistent with our goal to fit $E[Y(1) \mid X] - E[Y(0) \mid X]$ without making undue parametric assumptions.

BART for causal inference. It is straightforward to use BART to estimate the average treatment effect (ATE). First fit BART to the observed data (Y given Z and X). Next

make predictions for two datasets (Hill 2011). X is kept intact for both, however, in one all treatment values are set to 0, and in the other they are all set to 1. This allows BART to draw from the posterior distribution for $\mathrm{E}[Y(1) \mid X]$ and $\mathrm{E}[Y(0) \mid X]$ for each person, implying we can also obtain draws from $\mathrm{E}[Y(1) - Y(0) \mid X]$ for each person. These posterior distributions for individual-level treatment effects can then be aggregated to obtain posterior distributions of average treatment effects either for the full dataset or any subset thereof.

Adding the propensity score as a covariate. While the approach described in Hill (2011) has good properties across a variety of settings (Hill et al. 2011; Hill and Su 2013; Dorie et al. 2016; Kern et al. 2016; Wendling et al. 2018), recent work (Hahn et al. 2017) reveals situations where the performance of BART can be compromised due to regularization-induced confounding. While this is less of a concern in settings like the present one, in which covariates are outnumbered by observations and data are well-behaved, in general the "best practice" recommendation for using BART for causal inference is to guard against this potential source of bias. One approach to doing so, suggested by Hahn et al. (2017), is to include an estimate of the propensity score as a covariate. We used BART to fit a propensity score model (as described below) and included the estimate in response models.

Cross-validation to choose hyperparameters. BART tends to perform well using the default prior specification described by Chipman et al. (2007), but performance can sometimes be improved by choosing hyperparameters via cross-validation (Chipman et al. 2010). This is particularly important when using BART for non-continuous outcomes, a case for which off-the-shelf BART is currently not optimized (Dorie et al. 2016).[1]

Overlap. BART has certain advantages over propensity score approaches to evaluating overlap, which can be misled by covariates that are strongly predictive of the treatment but not are not strongly associated with the outcome. Therefore, in addition to checking overlap marginally for each covariate and for the propensity score, we also checked using a BART-specific approach as in Hill and Su (2013); results described in Section 3.

Causal inference with group structured data. The data have a multilevel structure. Treatment was assigned at the individual level, but students are grouped within schools. A primary goal was to decide whether and how to model this structure. During the workshop, we presented results from a fixed-effects specification. Ultimately, our preferred model for the response surface is the random-effects specification. However, robustness checks (below) reveal that this choice made little difference in overall results.[2]

1. Murray (2017) has derived models for a wide class of generalized linear model extensions to BART that are optimized for binary, count, and multiple category outcomes however these are not yet available in shareable software.
2. We used fixed effects for the propensity score model, since random effects with a binary response are not yet implemented in `dbarts`, or elsewhere in **R**, to our knowledge.

Overview of preferred modeling strategy. Our preferred modeling strategy proceeds as follows: 1) Fit a propensity score model with BART using all covariates and including school ID as a fixed effect, using cross-validation to choose hyperparameters (75 trees with k of 8). 2) Fit a response model on observed covariates and the estimated propensity score with BART including schools as random effects, using cross-validation to choose hyperparameters (350 trees with k = 1.5). For both fits we run 4 chains with 1000 iterations each (in addition to 500 burn-in iterations). Given the symmetry of the posterior distributions of interest, we report credible intervals based on normal approximations.

2. Results from analyses run for the workshop

During the workshop we discussed assumptions and addressed the questions posed.

2.1 Checking assumptions

Our first step was to check balance and overlap of covariates between treatment groups. We checked each covariate individually as well as the propensity score and found overwhelming support for both balance and overlap (see Figures A1 and A2 in Supplemental Appendix). We revisit overlap in Section 3 with more sophisticated diagnostics.

2.2 Goal 1: Intervention effectiveness

We addressed this question by using BART in the manner described above with a focus on estimating an average causal effect and associated uncertainty interval. The posterior distribution of this effect is reasonably symmetric, so we reported only an effect estimate (posterior mean) of 0.248 with a 95% credible interval of (0.227, 0.270). By this measure, we deem the intervention to be effective on average. Choice of grouping adjustment makes little difference to the estimate of the overall ATE leading to differences of less then .003 in posterior means and interval endpoints.

2.3 Goal 2: Moderation by specific covariates

We had several related strategies for exploring moderation. These capitalize on the fact that BART provides a posterior distribution of the causal effect for each observation. It is thus straightforward to examine the relationship between the expected effect for each person (represented by the mean of the corresponding posterior distribution) and any covariate of interest. We can do the same with respect to school-level effects and covariates. We present a few of the myriad methods for portraying these relationships.

The role of urbanicity. Before discussing our results for Goal 2, we address an important discovery made in our pursuit of that goal. While exploring the role of $X1$ ("fixed mindset")

and $X2$ ("achievement") as moderators we created scatterplots of individual treatment effect estimates (posterior means) versus $X1$ and $X2$. These plots revealed a group of schools with average treatment effects *substantially* lower than the rest, as displayed in the left-most plots of Figure 1.[3] Fitting a regression tree for the individual effects given all covariates (using the rpart package in **R**) easily identified the five-category, school-level covariate XC, or "urbanicity", as the culprit. Color-coding by urbanicity levels displays this visually. Posterior distributions of the treatment effect for each urbanicity level, displayed in the right-most plot of this figure, would have alerted us to this phenomenon as well. Of course, researchers do not typically check for moderation with respect to *all* covariates (and in fact are often discouraged from doing so out of fear of data snooping). Therefore, in the absence of a specific hypothesis about urbanicity, the substantial differences in treatment effects across its levels might have gone undetected with a more traditional test of moderation.

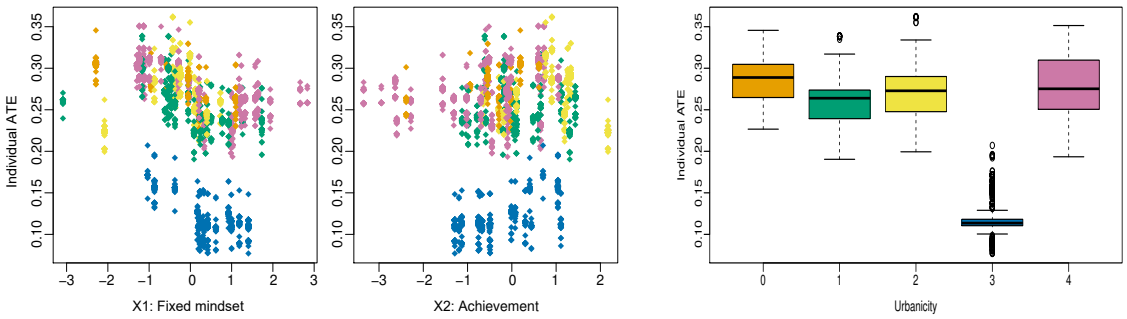


Figure 1: The role of urbanicity

Moderation by $X1$ and $X2$ Given the distinctive role that urbanicity plays in predicting school-level treatment effects, we opted to subtract the variation due to urbanicity from the school-level treatment effects. This was accomplished by centering the posterior mean individual ATEs on the average individual ATE within urbanicity category before computing school-level averages. In practice, this choice would be made in conjunction with the applied researcher, since it subtly changes the nature of the research question. In essence, we are now examining whether treatment effects vary across schools with the same urbanicity rating that differ in their average level of fixed mindset or their achievement.

For the workshop, we presented plots of the relationship between the school-level treatment effects and each of these potential moderators as lowess curves with uncertainty bounds as in Figure 2. These provide weak evidence of moderation by $X1$ with a trend towards smaller effects for schools that had higher levels of fixed mindset. Similarly there appears

3. Actually *first* we created lowess plots of these relationships. These masked this phenomenon! This is a testimony to the power of plotting your data!

to be some evidence for a positive association between school achievement and treatment effect. However, we weren't satisfied with using the default uncertainty bounds provided by ggplot for lowess. Our post-workshop analyses provide more clarity regarding these trends and associated uncertainty, but do not alter our overall conclusions.

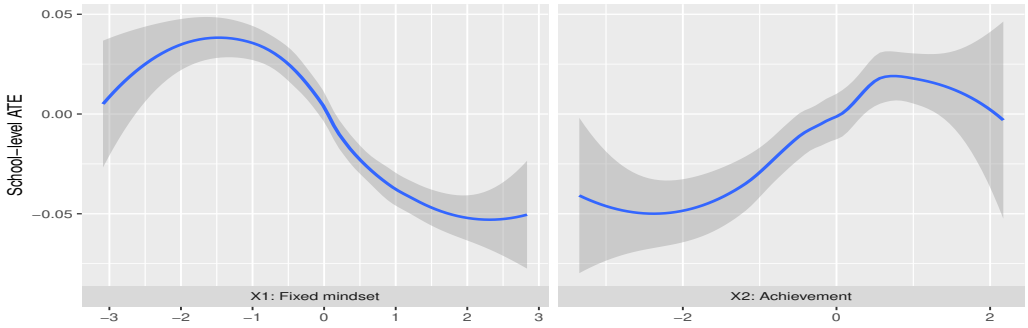


Figure 2: Lowess representations of $X1$ (left) and $X2$ (right) as moderators.

2.4 Goal 3: Moderation by other covariates

We presented moderation plots similar to those above for each of the continuous covariates ($X3$, $X4$, and $X5$); these are displayed in the Supplemental Appendix as Figure A3. We made more informative plots after the workshop; our conclusions did not change.

For binary covariates, we assessed moderation using the posterior distribution of the difference in ATE between groups. For "first-generation status" ($C3$), we observe a mean difference of -0.025 with 95% credible interval (-0.018, 0.060); the treatment effect is slightly (but not significantly) lower for first-generation students. There is no evidence of a difference in treatment effects by gender (difference estimate: -0.0095, 95% CI: (-0.44, 0.32)).

For multi-category factors, we began by simply examining side-by-side boxplots of the adjusted individual ATEs by level. As can be seen in Figure A4 in the Supplemental Appendix, there appears to be little evidence of a race effect. It is possible that there is a trend of increasing ATE as student expected success increases.

3. Post-workshop analyses

After the workshop we examined a few issues in more depth, as summarized here.

3.1 Re-examination of modeling choices

Our initial comparison of modeling strategies with regard to the grouping variable only considered differences in the overall ATE and corresponding uncertainty intervals. Given

the focus on moderation, we are interested in understanding whether the *individual* ATE estimates varied much based on this choice. Figure A6 in the Supplemental Appendix presents scatter plots of individual ATEs across all pairs of the three choices. There is little difference in estimates excluding the school variable and including it as a fixed effect. However, modeling the grouped structure using a random effect creates a noticeably wider distribution of individual-level effects.

We also examined more closely the impact of including the propensity score, by comparing our results to models without this feature. The estimate of the ATE in a model that excludes propensity score from the covariate set is 0.249 with associated credible interval (0.228, 0.270). This is almost identical to that of our preferred analyses. The correlation between posterior means of individual effects between these analyses is 0.896. We provide a more detailed comparison in the Supplemental Appendix.

3.2 Revisiting Moderation

We redid some of our original moderation analyses for several reasons. First, we found a way of graphically displaying our uncertainty about the relationship trends that is easier to interpret. Second, we wanted to more explicitly test the "Goldilocks" hypothesis posed by the research team. The analyses reported here are net of the impact of urbanicity on the treatment effects. Figure A7 in the Supplemental Appendix displays similar results without adjustment for urbanicity. These relationships are so dominated by the urbanicity-specific treatment effect differentials that they nearly all demonstrate a "reverse Goldilocks" pheonomenon. We start by discussing moderation by school-level achievement, $X2$, since the hypotheses regarding $X2$ are more complicated.

Moderation by $X2$. We explore the research questions about potential moderation by $X2$ in two ways. Our first approach partitions X2 into 3 subgroups using a recursive partitioning algorithm. Then treatment effects are averaged within subgroup to create draws from the posterior distribution of the average treatment effect for "low", "medium", and "high" values of $X2$. We can compare the "low" and "medium" subgroups or the "high" and "medium" subgroup by differencing the corresponding posterior distributions, as displayed on the left side of Figure 3. The posterior probability that the average treatment effect for medium subgroup is greater than for the low subgroup is 99.9%. However, the effect for the medium subgroup is *not* likely to be greater than the "high" subgroup - indeed, we find a posterior probability of 97.8% that the high subgroup has a *larger* average effect. This analysis does not provide evidence for the Goldilocks effect.

Second, we display on the right side of Figure 3 a scatter plot of school-level average treatment effects (net of urbanicity-specific means) versus $X2$ with a sample of quadratic fits to the posterior draws to illustrate our uncertainty about this fit. While limited by its

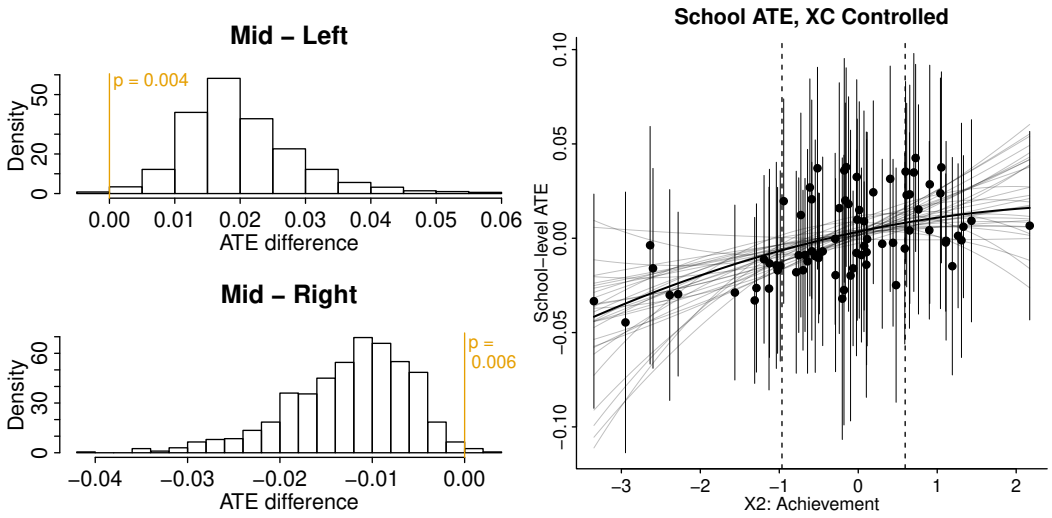


Figure 3: Left: Histograms of posterior distributions of differences in school average treatment effects between the medium and low $X2$ subgroups (top) and the high and medium subgroups (bottom). Right: Posterior distributions for school average treatment effects as a function of $X2$, after controlling for XC. Points are the posterior means of school average treatment effects and vertical lines show associated 95% posterior credible intervals. Curved lines show 30 samples from the posterior distribution of quadratic regressions fit to the school average effects (gray lines) and the posterior mean of all such regressions (black line).

simplistic parametric form, examining the coefficient of the squared term offers a straightforward test for a rise-then-fall relationship. The posterior mean indicates a slight Goldilocks effect, however the posterior uncertainty in the square-term coefficient is consistent with no effect. There is only an approximate 68.7% posterior probability of this term being negative.

Even cursory visual inspection of Figure 3 discounts the alternative hypothesis that the higher school-level achievement is associated with smaller treatment effects.

Moderation by $X1$. The relationship between school-level treatment effects and fixed mindset, $X1$, is decreasing without strong evidence of quadratic curvature, as displayed in Figure 4. The probability that the linear part of this decreasing trend is less then zero is approximately 96.6%, and the probability that the quadratic part is less than zero is 75.8%.

Moderation by the other school-level continuous covariates. In Figure 5 we display moderation plots similar to those in the previous section for the remaining continuous school-level variables: "minority composition" ($X3$), "poverty concentration" ($X4$), and "school size" ($X5$). The posterior probabilities that the linear terms are negative are ap-

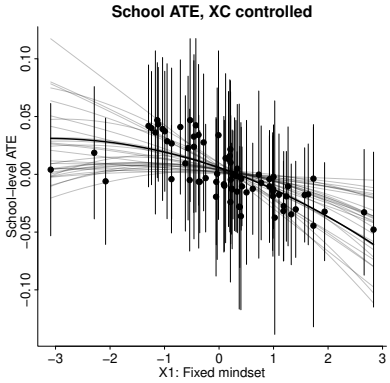


Figure 4: Posterior distributions for school average treatment effects as a function of X1, after controlling for XC. All else is as described in Figure 3.

proximately 84.0%, 73.0%, and 10.6% respectively, while the corresponding probabilities for the quadratic terms are 90.8%, 91.4%, and 6.3%. This provides some evidence of a Goldilocks effect for poverty concentration but nothing earthshattering.

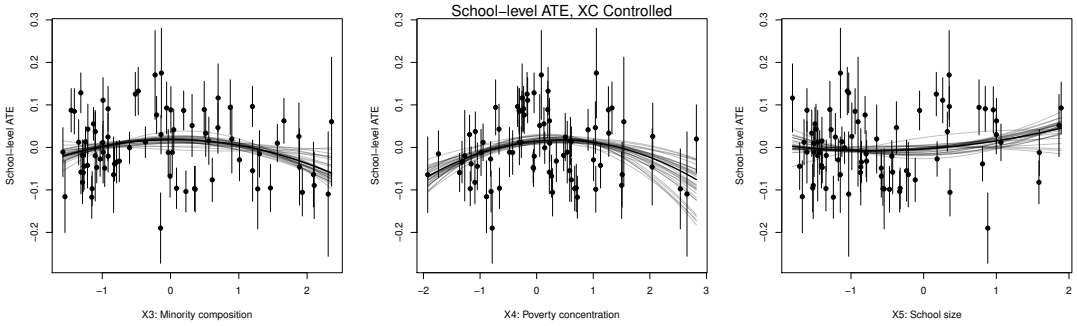


Figure 5: Posterior distributions for school average treatment effects as a function of $X3$, $X4$, and $X5$, net of XC. All else is as described in Figure 3.

Moderation by Student Expected Success: A closer look. We return to examine moderation by $S3$, student expected success, because our initial results provided some evidence for a moderated effect but we performed no formal tests.[4] Figure 6 displays a line plot connecting the posterior means of the ordered categories along with corresponding credible intervals. The right plot tests whether levels 6 and 7 have larger effects than those below; there is moderate support (91% probability) for this hypothesis.

4. The Supplemental Appendix provides a somewhat similar reanalysis for the race variable.

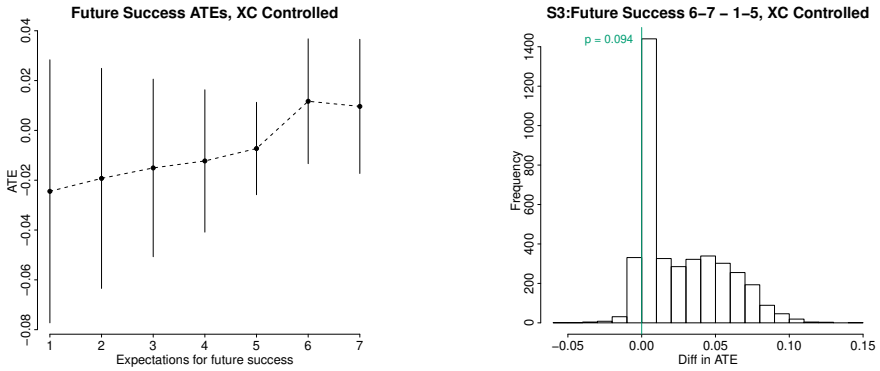


Figure 6: Left: Means and 95% credible intervals of posterior distributions (impact of XC removed) for each level of ordered categorical variable $S3$ presented as a line plot. Right: Posterior distribution for the difference in mean effects between the two top levels of $S3$ and the rest. About 91% of this distribution lies above zero.

3.3 More formal checks of assumptions

BART has already been incorporated into diagnostic frameworks to examine the plausibility of two crucial causal assumptions: ignorability and overlap.

Dorie et al. (2016) demonstrate how BART can be incorporated into a sensitivity analysis framework to help researchers to understand under what conditions their results might be sensitive to unobserved confounding. This approach is available in the treatSens package on CRAN as the function `treatSens.BART`. The results from this sensitivity analysis are displayed in Figure A10 in the Supplementary Appendix. This plot reveals that the level of unobserved confounding would need to be extremely strong in order to remove the estimated positive effect. This sensitivity analysis strongly supports our assumption of ignorability.[5]

Our covariate-by-covariate examination of overlap in the previous section (results displayed in the Supplemental Appendix) provided strong evidence in support of marginal balance and overlap. In the Supplemental Appendix we present two additional looks at the issue. The first, displayed on the left side of Figure A9, is a scatter plot of the joint distribution (estimated posterior means of) $Y(0)$ and $Y(1)$ for each observation for both treated (red) and control (blue) observations that suggests strong overlap.

To investigate local overlap (as suggested in the workshop discussion) we calculated an overlap statistic recommended by Hill and Su (2013). For each person we calculate the ratio of the variance of the posterior distribution of their counterfactual outcome relative to the variance of the posterior distribution for their factual outcome. The distribution of these ratios is displayed in Figure 4 A9. We gauge the extremity of any such ratio relative to a

5. The treatSens package does not yet accommodate random effects so was run with fixed effects.

Chi-squared distribution; the 10% cutoff would be about 2.7; no ratios even come close to this threshold, suggesting that overlap is present locally as well as marginally.

4. Discussion

Several features make BART a powerful tool for causal inference. The sum-of-trees model flexibly fits even highly non-linear response surfaces. The Bayesian inferential framework allows us to easily quantify our uncertainty not only about the average treatment effect and individual-level treatment effects but also any functions of the potential outcomes (all without re-using our data). The recent extensions that accommodate varying treatment effects extend the applicability of this tool to simple multilevel data structures.

We used BART to address the questions posed and found strong evidence of a large positive average effect of the intervention (the "effect size" is about .4 and the credible interval has near zero probability of covering 0). Urbanicity strongly moderates this treatment effect therefore we addressed the other questions after adjusting for this.[6] Net of urbanicity, we find some evidence of moderation at the school level: the school-level mean of students' fixed mindsets, $X1$, is moderately negatively associated with the size of effect; achievement, $X2$, is moderately positively associated. Of all student-level variables there is most support for moderation by student expected success, $S3$.

Our results are predicated on satisfying several assumptions—ignorability, overlap, etc.—that in some situations can be heroic. The BART extensions that easily allow examination of the evidence for and implications of these assumptions add credibility to our analyses. We found strong support that these assumptions were satisfied.

References

Chipman, H., George, E., and McCulloch, R. (2007). Bayesian ensemble learning. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298.

Dorie, V., Carnegie, N. B., Harada, M., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, 35(20):3453–70.

6. In a "real-life" situation where we could interact with the applied researcher we might make a different choice based on their understanding of the theoretical questions of primary interest. However we were making decisions in the absence of the availability of such an interaction.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2018). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, accepted with discussion.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *ArXiv e-prints*.

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hill, J. and Su, Y.-S. (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.*, 7(3):1386–1420. Available from: http://dx.doi.org/10.1214/13-AOAS630.

Hill, J. L., Weiss, C., and Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, 46:477–513.

Kern, H. L., Stuart, E. A., Hill, J. L., and Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target samples. *Journal of Research in Educational Effectiveness*, 9:103–127.

Murray, J. S. (2017). Log-Linear Bayesian Additive Regression Trees for Categorical and Count Responses. *ArXiv e-prints*.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*.

Appendix A. Supplemental Appendix to Carnegie et al. Discussion

A.1 Additional workshop analyses

Overlap plots We examined the overlap and balance of each of the covariates marginally through a variety of plots; see Figure A1 and Figure A2. These demonstrated a high degree of both balance and overlap.

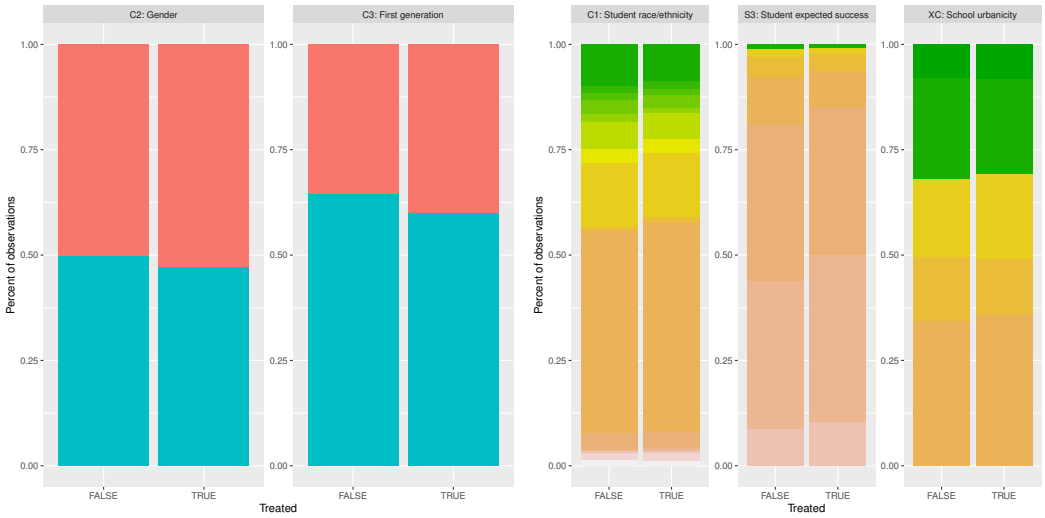


Figure A1: Overlap in student-level binary variables (left) and multiple-level categorical variables (right)

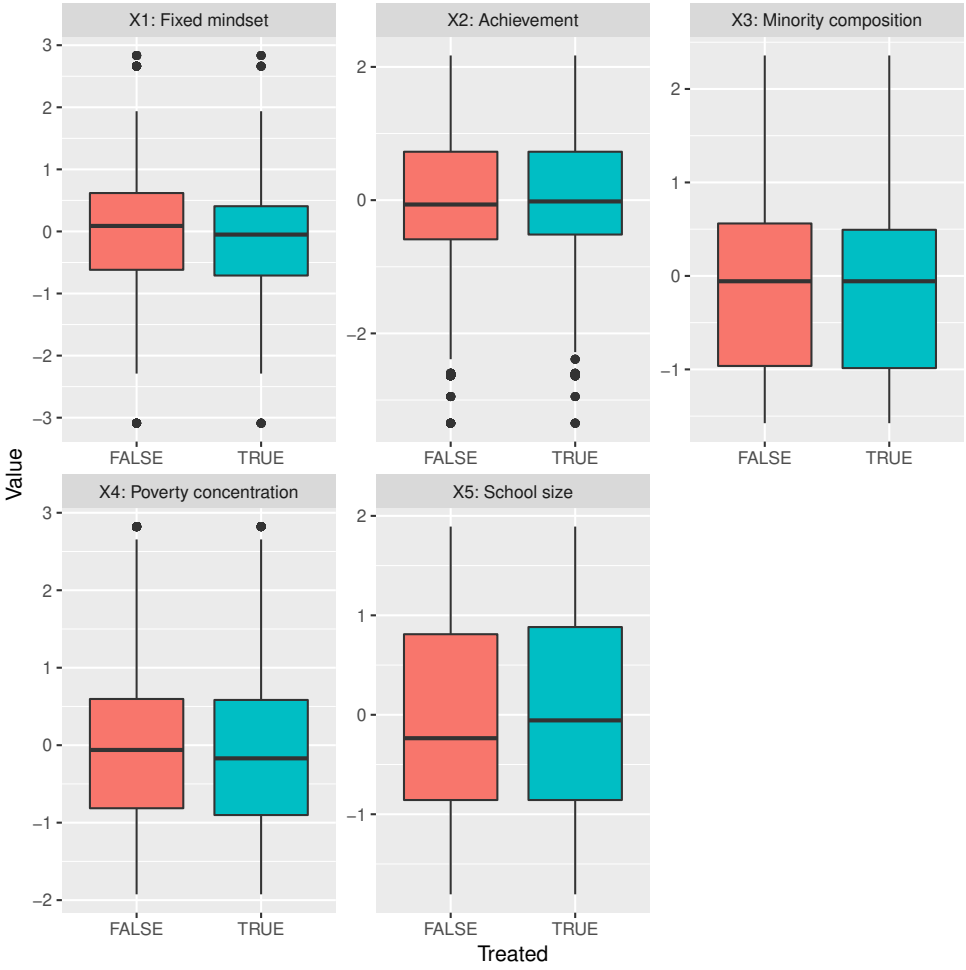


Figure A2: Overlap in school-level continuous variables

Effect modification by continuous variables: lowess plots For the workshop, we presented plots of the relationship between the school-level treatment effects and each of the potential moderators as lowess curves with uncertainty bounds. Figure A3 gives the resulting plots for school-level continuous covariates X3 through X5.

Figure A3: Lowess representations of X3 (left) through X5 (right) as moderators.

Effect modification by categorical variables: boxplots For categorical variables, we used simple side-by-side boxplots to evaluate potential effec modification. There was little evidence of an effect of race using this method, but some suggestion of an increasing effect with student expected success (S3). In particular it appears that the treatment effect is larger for students whos expected success is greater than 5.

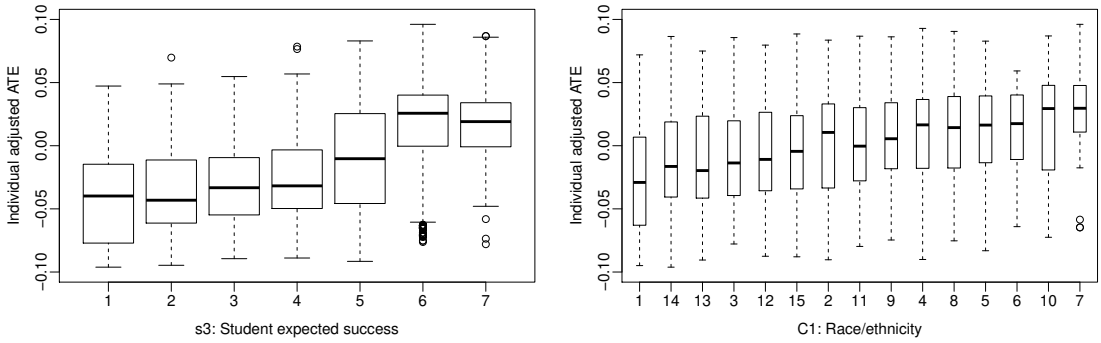


Figure A4: Boxplots of adjusted individual ATE by of S3 (left) and C1 (right) as moderators.

A.2 Additional results from post-workshop analyses

The impact of group-level modeling strategies The plot in A6 displays a scatter plot of individual ATE's that displays how they vary across adjustment methodologies: no

adjustment (that is, including school ID as a continuous covariate so that BART is forced to make splits based on difference between continguous categories), fixed effects, and random effects.

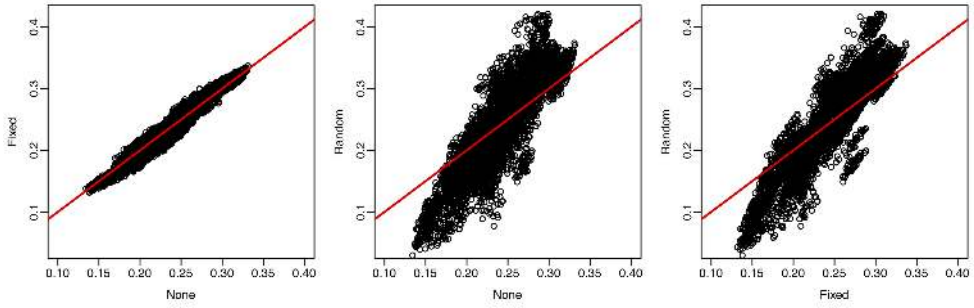


Figure A5: Scatterplots of individual ATE estimates across adjustment methodologies.

Impact of adding the propensity score as a covariate We discuss in the main text the high degree of correspondence between the individual-level treatment effect estimates produced using an estimation strategy that includes the estimated propensity score as a covariate versus one that excludes it. However a scatter plot of the two sets of estimates reveals an interesting feature which is that these estimates appear to come from a mixture of two subpopulations. When we predict the random effect estimates using a regression tree $S3$ emerges as by far the strongest predictor. Highlighting the pattern we saw in Figure A4.

Treatment Effect Modification when urbanicity has not been netted out. The plots in the main test display results examining associations between covariate and treatment effects. We felt it was also important to examine how different these results might be if we had decided not to net out urbanicity (XC). Figure A7 shows school average treatment effects as a function of X2 with levels of XC highlighted by color. Urbanicity category 4 is markedly below the others and it complicates one of the primary objectives of this exercise: characterizing the moderating effect of X2. Not only are the hypotheses that X2 has a "Goldilocks" impact on the treatment effect or that it steadily decreases effectiveness ruled out, if urbanicity is not controlled for one can reach an opposite conclusion - that of least effect in the middle range. Consequently, all future analyses are done by controlling for urbanicity and subtracting out the level average effects.

Moderation by Race: A closer look We revisited moderation by race after the workshop to implement some more formal tests. Figure A8 shows the racial average treatment effects after controlling for urbanicity (XC). It provides some evidence of racial moderation of the treatment effect, however many of the effects are consistent across race categories.

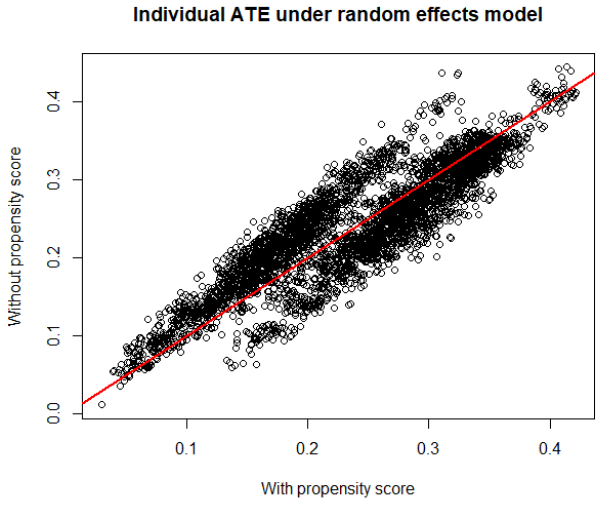


Figure A6: Scatterplot of individual ATE estimates using random effects model with and without propensity score included as a predictor.

The highest and lowest race average treatment effects are estimated with a considerable degree of uncertainty due to their sample sizes, however the posterior distribution of the difference between the highest and lowest racial averages - based on their posterior means - yields a borderline "statistically significant" difference. The distribution over this difference has 5.2% probability assigned to negative values, so that a one-sided posterior credible interval would just barely include 0. However this contrast was chosen after looking at the plots and without a clear hypothesis about "race level 11" as a specific moderator. Therefore we see such analyses as exploratory.

Diagnostics that assess plausibility of the ignorability and overlap assumptions. Figure A9 displays evidence regarding the overlap based on BART output as described in the main text.

Sensitivity to unobserved confounding. We see that the amount of confounding necessary to substantively change our results would be quite extreme and certainly far exceeds the current levels of associations with observed covariates.

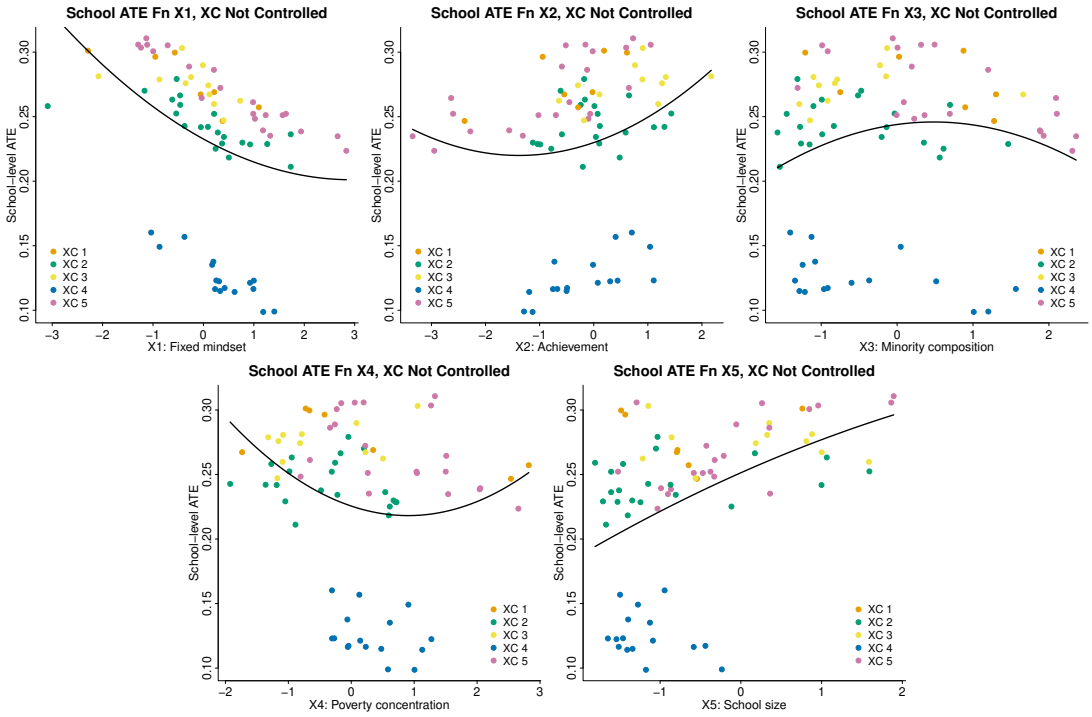


Figure A7: Posterior distributions for school average treatment effects as a function of, left-to-right/top-to-bottom, X1, X2, X3, X4, X5. The points are the posterior means in each school while the curved line is the posterior mean of quadratic regressions fit to the school average treatment effects.

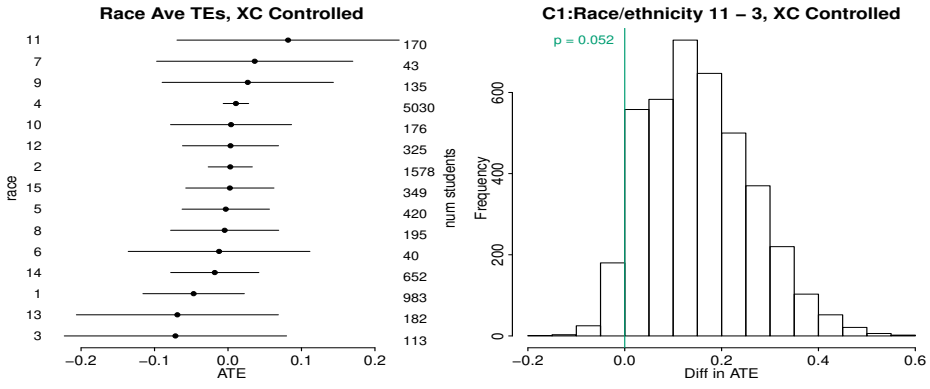


Figure A8: Left: Posterior means and 95% credible intervals of average treatment effects for each race category after controlling for XC. Right: Histogram of posterior samples for the difference between the highest and lowest treatment effect races.

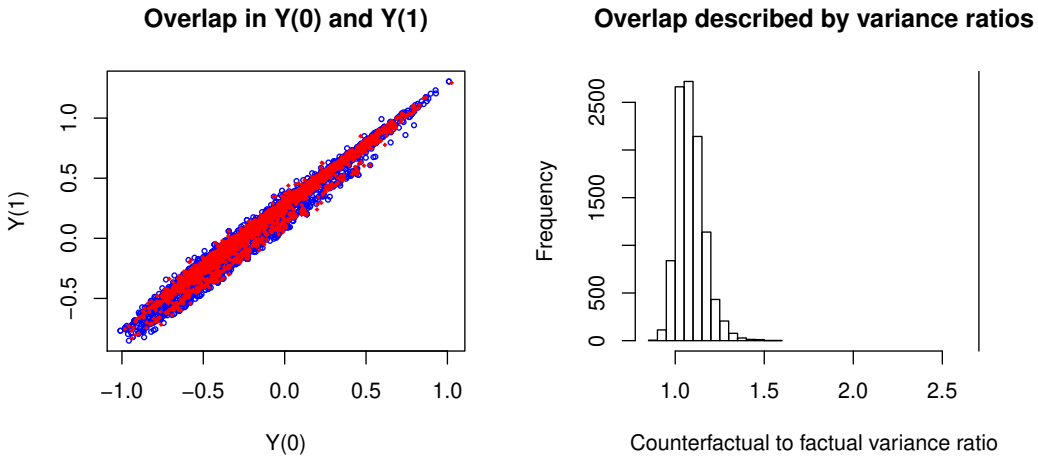


Figure A9: Left: Overlap across treatment (red) and control (blue) groups with regard to distribution of Y(0) and Y(1). Right: Distribution of variance ratios for counterfactual versus factual outcomes.

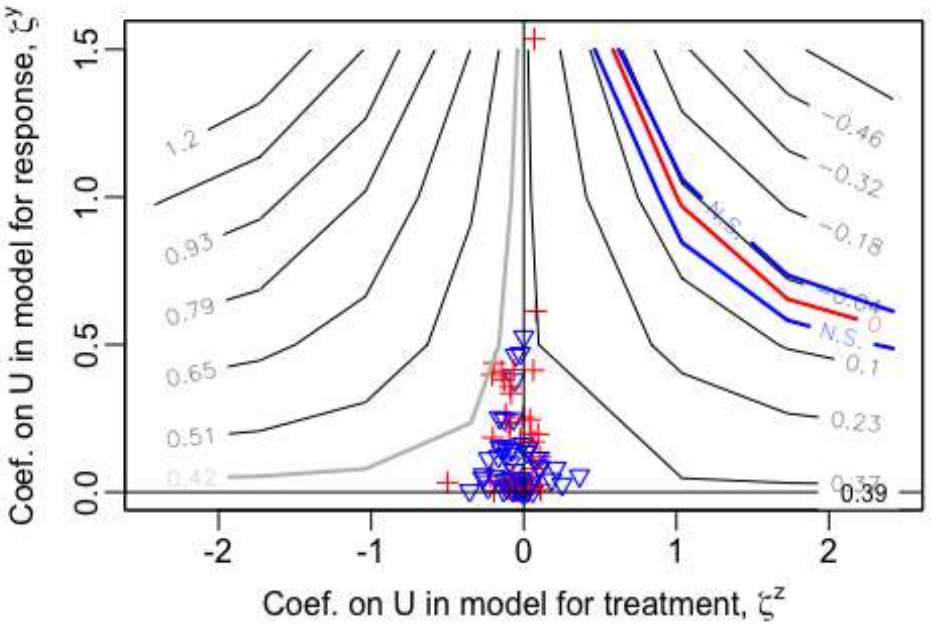


Figure A10: Sensitivity to unobserved confounding.