Using Hierarchical Models to Estimate Heterogeneous Effects

Joshua Alley
Assistant Professor
Baylor University
Joshua_Alley@baylor.edu

October 1, 2025

Abstract

This note describes why, when, and how to use Bayesian hierarchical models to estimate heterogeneous effects. While an ample literature suggests that hierarchical models provide helpful regularization and information about effect variation, political scientists rarely use them to estimate heterogeneous effects. Doing so is simple, however. To start, identify potential sources of heterogeneity. Then, fit a hierarchical model with two linked regressions, one connecting treatment with the outcome, and another that models the treatment effects with potential sources of heterogeneity partially pools individuals. This captures systematic group-level variation and random individual variation in heterogeneous effects, encompasses the diversity of interactions in theories, and fits commonly used modeling frameworks. Hierarchical modeling is more flexible than linear interactions and reduces the risk of underpowered subgroup comparisons. It also provides a more interpretable framework for testing theories than machine-learning tools. I document these claims with a simulation analysis and extension of a published study. Researchers can thus use hierarchical models alongside other approaches to understand heterogeneous effects for scholarship and policy.

1 Introduction

Whether in observational or experimental studies, every independent variable social scientists examine impacts some units differently than others. Common estimands aggregate heterogeneous effects.[1] Such average effects are useful, but they often obscure interesting and important variation.

Understanding heterogeneous effects is essential for policy and scholarship. Estimating heterogeneity allows scholars to clarify when their independent variable most or least impacts some outcome. Policymakers can maximize the impact of finite resources with targeted interventions, for example by providing job training to individuals who are more likely to benefit.

This paper explains why, when and how to use hierarchical models to estimate heterogeneous effects. A large statistics literature suggests that Bayesian hierarchical models are a useful tool for heterogeneous effects estimation (e.g., Feller and Gelman (2015); McElreath (2016); Dorie et al. (2022)). Political scientists tend to rely on interactions or machine learning tools instead, however. For instance, of the three applied political science citations of Feller and Gelman (2015), only Marquardt (2022) models treatment effects.

This oversight matters because there are few tools that are well-suited to test the proliferation of conditional arguments in the social science. Social scientists often propose conditional theories (Clark and Golder, 2023) and are interested in how different people respond to the same stimulus for normative or policy reasons. Many theories proposing a singe modifier for the same relationship and interest in diverse subgroups suggest that multiple modifiers are the rule, not the exception. For example, scholarship on audience costs has considered how foreign policy dispositions (Kertzer and Brutger, 2016), partisanship (Levendusky and Horowitz, 2012), gender (Barnhart et al., 2020; Schwartz and Blair, 2020) and policy preferences (Chaudoin, 2014) modify individual reactions to a leader backing down from a threat. Such a mul-

[1] For instance, Abramson, Koçak and Magazinnik (2022) note that the average marginal component effect (AMCE) of conjoint experiments gives more weight to intense preferences.

tiplicity of theoretically informed modifiers complicates empirical testing, however.

Scholars cannot ignore heterogeneity, but the most common tools either increase the risk of spurious results or are hard to interpret and use. Interaction terms and subgroup analysis are perhaps the most common tool, but can mislead. Simple interactions and subgroup analyses are ubiquitous because they are relatively easy to interpret, but these have serious power concerns. Many political science analyses have low power even to detect main effects (Arel-Bundock et al., 2022). Adequate power for estimates of even a single interaction can require significantly more data (Gelman, 2018), which may be prohibitively expensive or impossible. As a result, statistically significant heterogeneous effect estimates may be far too large— the result of noise in the data, not systematic differences. This problem is partially responsible for widespread issues replicating findings based on interactions (Simmons, Nelson and Simonsohn, 2011).

Scaling up the number of modifiers is therefore something that theory suggests scholars should do, but doing so with interactions is not easy. The interpretation benefits of parametric interactions diminish as the number of modifiers increase- the different terms in interactions of three or four variables can be difficult to parse. It also further raises the risk of spurious inferences due to power concerns and picking up noise in ever more finely sliced subgroups.

Given multiple sources of heterogeneity, machine-learning tools such as random forests (Green and Kern, 2012; Wager and Athey, 2018), support vector machines (Imai and Ratkovic, 2013), and ensemble methods (Grimmer, Messing and Westwood, 2017; Künzel et al., 2019; Dorie et al., 2022) are more likely to avoid over-fitting. These machine learning algorithms usually have some regularization component and can discover complex patterns and high-dimensional variation across multiple modifiers.[2] These tools can be difficult to interpret and implement, especially in smaller social science datasets. A lack of interpretability is especially problematic for testing the relative weight of multiple modifiers.

[2]Blackwell and Olson (2022) describe a lasso approach to interactions that sits between machine learning and linear regressions.

The hierarchical strategy I propose here addresses the problems of interactions with regularization, and issues of interpretation in machine learning. I do this by showing how scholars can use two connected regressions to estimate theoretically informed models of multiple heterogeneous effects. Using hierarchical models is more flexible than standard interactions but easier to implement and interpret than machine learning approaches. It preserves a straightforward structure while accommodating more factors and ameliorating the downsides of subgroup analysis. This facilitates argument testing. The main downside is that unlike machine learning, the hierarchical approach lacks the flexibility to discover high-dimensional heterogeneity. Hierarchical modeling therefore works best when theory indicates more than two modifying factors and there is less emphasis on discovery.[3]

There are two key steps when theory and data make using hierarchical models worthwhile. First, researchers should identify potential modifiers of a treatment, and determine the most theoretically appropriate model of treatment effects. Second, they should take that model of treatment effects and connect it to a model with individual effects linking a treatment and outcome. Modeling heterogeneous effects in this way produces interpretable results, which facilitates argument testing. It also allows researchers to compare different sources of heterogeneous effects and describe how much an effect varies.

While frequentist estimation of hierarchical models is possible, Bayesian estimation is straightforward and more informative. Bayesian estimation provides crucial information by connecting parameters through common prior distributions, thereby regularizing estimates and propagating uncertainty. Working with posterior distributions also gives researchers more flexibility to describe how and when effects vary. While computation and coding were once a barrier to employing Bayesian methods, fitting a wide range of hierarchical models is straightforward

[3]Goplerud (2021) introduces a model that uses Bayesian structured sparsity to estimate which group coefficients are similar and which are different. In this approach, researchers use theory to inform potential groups, but the data determines common estimates for groups.

with the brms package in R (Bürkner, 2017).[4]

In the remainder of this paper, I describe how and when to estimate hierarchical models of heterogeneous effects. I then employ a simulation study to compare OLS and hierarchical estimates of individual treatment effects under different conditions. Finally, I demonstrate the process by analyzing a study of how military alliances shape public support for war by Tomz and Weeks (2021). The reanalysis reveals that alliances increase support for intervention most among white men who support international engagement but are otherwise skeptical of using force.

2 Hierarchical Modeling of Heterogeneous Effects

There are two steps in hierarchical models of heterogeneous effects. First, researchers must identify potential sources of heterogeneity, and think about the right model of how those factors shape heterogeneity. This will also depend on what variation is most important and interesting. Theory, policy concerns, or normative factors are all possible motivations.

This first step determines what heterogeneous effects a researcher estimates. It is analogous to researchers thinking through their regression specifications— the same sort of care should go into the sources of heterogeneity. Researchers need to define what variation is most important, links heterogeneous effects to theory, and structures modeling.[5] Not thinking carefully about sources of heterogeneity will obfuscate results and can hinder model fitting.

There are three general approaches to defining groups. First, researchers can set groups using combinations of other treatments, especially when an intervention has several dimensions but theory emphasizes one of them. The experimental design determines modifiers, and the model estimates heterogeneous treatment effects. If researchers want to know how different issues shape the impact of elite foreign policy cues (Guisinger and Saunders, 2017), they

[4] I provide example code in this note and the appendix.
[5] It also facilitates pre-registration when applicable.

could include indicators of issues, for instance. Hierarchical estimators for topic-sampling experiments estimate how a treatment effect varies across different topics (Clifford and Rainey, 2023).

Researchers sometime use fully crossed regression interactions to estimate the impact of a treatment across experimental strata, but this approach risks spurious results by functionally estimating subgroup results, most of which do not have adequate power.

A second approach uses unit, demographic and contextual factors to estimate effect heterogeneity. Here, researchers examine what factors within or around units shape their response to an independent variable. For example, one could use a mix of individual and state-level factors to predict divergent consequences of a survey experiment treatment. Other use cases include estimating how different geographic units respond to an intervention.[6]

For example, Alley (2021) uses alliance characteristics to examine when alliance membership increases or decreases military spending. He models the impact of alliance participation as a function of treaty depth, partner democracy, conditions on military support, issue linkages, democratic membership, foreign policy concessions and other factors. All of these variables are potential sources of credibility or confounding factors. Democratic alliances have higher depth, so this model of heterogeneity accounts for potential confounding, and finds that after accounting for depth, democracy does not impact the relationship between alliances and defense spending.

Third, researchers might use hierarchical models to address specific policy concerns. Policy analysts often want to know how an intervention impacts a specific population. Researchers might want to know if a job-training program improves employment outcomes for black women in the South, for instance. To do this, a researcher might specify a heterogeneous effects model with interactions of race, gender and region, plus additional controls or other factors.

[6]Extrapolation to a representative sample for such units might require poststratification.

After defining groups, the second step is fitting a hierarchical model that links a model of the outcome with a model of heterogeneity. Essentially, researchers model the outcome and the process that produces heterogeneous treatments. The model employs two connected regressions. One regression deals with the outcome. The other regression models the treatment effects.[7]

I now briefly describe the generic hierarchical model. For ease of exposition, consider making between-unit comparisons based on an experimental treatment. Start with N units indexed by i, some of which receive a binary treatment T. Assume that the outcome variable y is normally distributed with mean μ_i and standard deviation σ.[8]

The outcome for each unit depends on an overall intercept, an optional matrix of control variables $\mathbf{X}$, and a set of individual treatment effects λ_i. When T is binary, estimated λ parameters for untreated units have no impact on the outcome. For a continuous treatment, the impact of of treatment will depend on the product of T_i and λ.

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma) && \text{(Likelihood)} \\
\mu_i &= \alpha + \lambda_i T + \mathbf{X}\beta && \text{(Outcome Equation)} \\
\lambda_i &= \theta_i + \mathbf{Z}\gamma && \text{(Heterogeneity Equation)} \\
\theta_i &\sim N(\mu_\theta, \sigma_\theta) && \text{(Indiviudal Varying Intercepts)}
\end{aligned}
$$

The heterogeneity equation then models those individual treatment effects with a systematic and random component. The systematic component is a matrix of predictors $\mathbf{Z}$ and associated parameters γ. Z can mirror any regression specification researchers might use for

[7] If other units such as states define the groups, rather than combinations of modifying variables, then adding group-level predictors is essential. For examine, in a model where an effect varies by state, adding state-level variables like ideology, population and GDP would avoid partially pooling small groups too far towards the overall mean.

[8] Researchers can and should use binary, categorical and other outcome likelihoods.

an outcome; linear combinations of variables, interactions, or other terms. Researchers might use interactions to capture processes where combinations of modifiers produce non-additive jumps in heterogeneity.

The random component of the heterogeneity equation is a series of individual-specific varying intercepts θ_i. These are critical, because they capture individual-specific deviations from the systematic trends expressed in the design matrix $\mathbf{Z}$. Outliers that otherwise might bias the λ estimates are partially pooled back towards the overall mean μ_θ.[9]

The above model can be fit with Bayesian or frequentist methods, but Bayesian estimation offers important advantages. First, it is more flexible, and including prior information can facilitate model fitting and convergence. Putting priors on the α, β, and γ parameters is especially helpful. Priors also help regularize estimates by pulling extreme groups towards the overall mean. Working with posterior distributions also also provides a wealth of information about effect heterogeneity and propagates uncertainty.

In interpreting these estimates, researchers should leverage the full range of information from the different parameters. First, the λ posteriors give the impact of the treatment on each individual, and are the core quantity of interest. All λs reflect a systematic component from the predictors in $\mathbf{Z}\gamma$ and a random variations from θ. γ parameters can, depending on the regression, be interpreted as the impact of a change in a modifier on the treatment effects. For example, a γ of .1 on a binary modifier means that λ is .1 higher in expectation when the modifier is one, and .1 lower when it is zero. σ_θ thus measures the extent of individual variation that is outside the systematic regression. Other techniques such as interactions in OLS with robust standard errors provide less information.

[9] In brms, using non-linear syntax can express a model with a treatment, two controls, and three modifiers as: y lambda * treat + control1 + control2, lambda mod1 + mod2 + mod3, nl = TRUE

3 When to Use Hierarchical Models

In deciding whether to use a hierarchical model, researchers must weigh specific advantages and disadvantages. In general, estimating heterogeneous effects in this way has three advantages. First, researchers can make detailed inferences about heterogeneous effects in an interpretable framework. This helps examine theories that predict how an effect varies and compare sources of variation.[10] Partial pooling also facilitates reasonable estimates for small groups by sharing information across groups and incorporating predictors in the heterogeneous effects equation. Finally, this approach will be faster than machine learning approaches for many datasets as well as easier to use in small datasets.

Like all methods, the hierarchical approach has downsides, some of which can be ameliorated with modifications, while others should lead researchers to use different tools. First, defining many small groups, perhaps by grouping based on a continuous variable, will likely lead to model fitting problems. If using continuous variables hinders model convergence, researchers can bin continuous variables. In general, fitting a model with many small groups will make it harder to fit a hierarchical model.

Furthermore, hierarchical models can show general trends, but will not make powerful comparisons between every group. Researchers who want to compare specific groups may lack empirical leverage, especially for small groups. This downside can also apply to other methods, however.

With these considerations in mind, when should researchers use hierarchical models in place of interactions? If only one factor modifies an effect, interactions are best, as the extra information hierarchical models provide is less valuable. Especially if the groups are of similar size and there is one grouping factor, regularization and OLS will give similar answers.

With two or more modifiers, hierarchical models begin to add value beyond. Interpreting

[10] Rescaling variables in the heterogeneous effects equation can aid model fitting and coefficient comparisons (Gelman, 2008).

triple interactions between a variable and two modifiers is challenging. The advantages of hierarchical modeling increase with the number of modifiers, until additional modifiers create small groups that complicate model fitting. The thresholds where the number of modifiers becomes an issue for hierarchical modeling depends on the data, as larger datasets can support more groups.

The relative use cases of hierarchical models and machine learning are different. Unlike machine learning approaches, hierarchical models will not discover high-dimensional interactions. Researchers can add flexibility with additional interactions or non-linear specifications in either level of the model, but this requires a priori specification. Therefore, if researchers want to focus on flexible discovery, not testing an argument with multiple sources of treatment heterogeneity, they should rely more on machine-learning.

In summary, researchers should continue to use interactions for single modifiers and machine learning to discover complex interactions. Hierarchical modeling works well when there are two or more modifiers and researchers have adequate data to support an informative model. Table 1 summarizes some relevant characteristics of hierarchical, interaction and machine learning approaches to heterogeneous effects. Hierarchical modeling is thus an intermediate tool between interactions and machine-learning, where researchers need more flexibility than interactions but are not willing or able to tackle the computational and interpretation challenges of machine learning.

4 Performance on Simulated Data

To assess how hierarchical models compare to interactions in OLS models, I first assess their performance on simulated data. The first simulation manipulates variation in the treatment effects and outcome, while leaving the number of groups fixed. Adjusting these factors changes the potential benefits of regularization. To second simulation varies the same aspects of the

	Hierarchical Models	Interactions/Subgroup	Machine Learning
Factors	Two or more	One or two	Many
Sample Size	Conditional on number of factors	Medium to large, depending on main effect size	Large
Complexity	Medium	Low	High
Computational Cost	Medium or High	Low	High
Interpretability	High	High	Low
Modifiers	Specified	Specified	Discovered or Specified

Table 1. *Key characteristics of different approaches to estimating heterogeneous effects.*

data-generating process and also manipulates the number of groups. Both simulations use 1,000 observations, so increasing or decreasing the number of groups changes the number of data points in each group.

The first simulation defines groups based on fully crossed interactions of five binary variables. I base the group treatment effects on interactions between the treatment and a dummy indicator for each group. I simulate the interaction coefficients by drawing them from a normal distribution with a mean of 0 and standard deviation of .05, .25 or .75. Increasing the standard deviation adds greater variation to the group-level effects. This captures the degree of heterogeneity in the relationship.

The interactions predict the mean of the outcome variable, μ_y, which has a standard deviation of .05, .25, or .75. The key treatment effect is the difference in μ_y between treatment and control observations within each group. I use differences in mu_y as the true value because this is the systematic component of the simulation- the observed outcome y has a random component.

I then fit two models. First, I fit an OLS model that uses interactions of the grouping dummies with the treatment variable. Next, I fit a hierarchical model that estimates varying intercepts and treatment slopes across groups.

There are several potential metrics for comparing model performance (Hopkins et al., 2024). One is bias; the gap between the estimated treatment effects in each group and the true treatment effects. Hierarchical models are sometimes biased on average, however (Clifford and Rainey, 2024), and the question is whether reduced variance in the estimates offsets bias enough to improve the estimates. To assess that, I compare the root mean squared error of the group treatment effect estimates for each model. Improvements in the root mean squared error suggest that that reduced variance overcomes any bias in the hierarchical estimates, improving overall model performance (Hopkins et al., 2024, pg. 39). In the appendix, I show some other performance statistics, including...

Figure 1 summarizes the results of the first simulation. This figure plots the difference in root mean squared error between the hierarchical and OLS estimators. Negative values favor the hierarchical model, while positive values mean OLS performs better.

When the outcome has low variance, OLS and hierarchical models perform similarly. This implies that regularization is most important when the outcome is noisy. Hierarchical models offer slight improvements at the next highest level of outcome variance, regardless of how much variance there is in the coefficients themselves. Hierarchical modeling produces the largest gains at the highest levels of outcome variance, especially when the coefficients vary less than the outcome. As is often the case, the regularization benefits of hierarchical modeling are greatest when there is substantial noise in the data-generating process. OLS estimates of heterogeneous effects are more sensitive to random variation in the outcome.

In the second simulation, I define between 2 and 6 groups in a dataset with 1,000 observations. Each group has the same number of data points, so as the number of groups increases, the number of observations in each group falls. More and smaller groups will likely benefit more from regularization, improving the performance of hierarchical modeling.

Figure 2 summarizes the results of the second simulation. Again, this figure plots the difference in root mean squared error between the hierarchical and OLS estimators, and negative

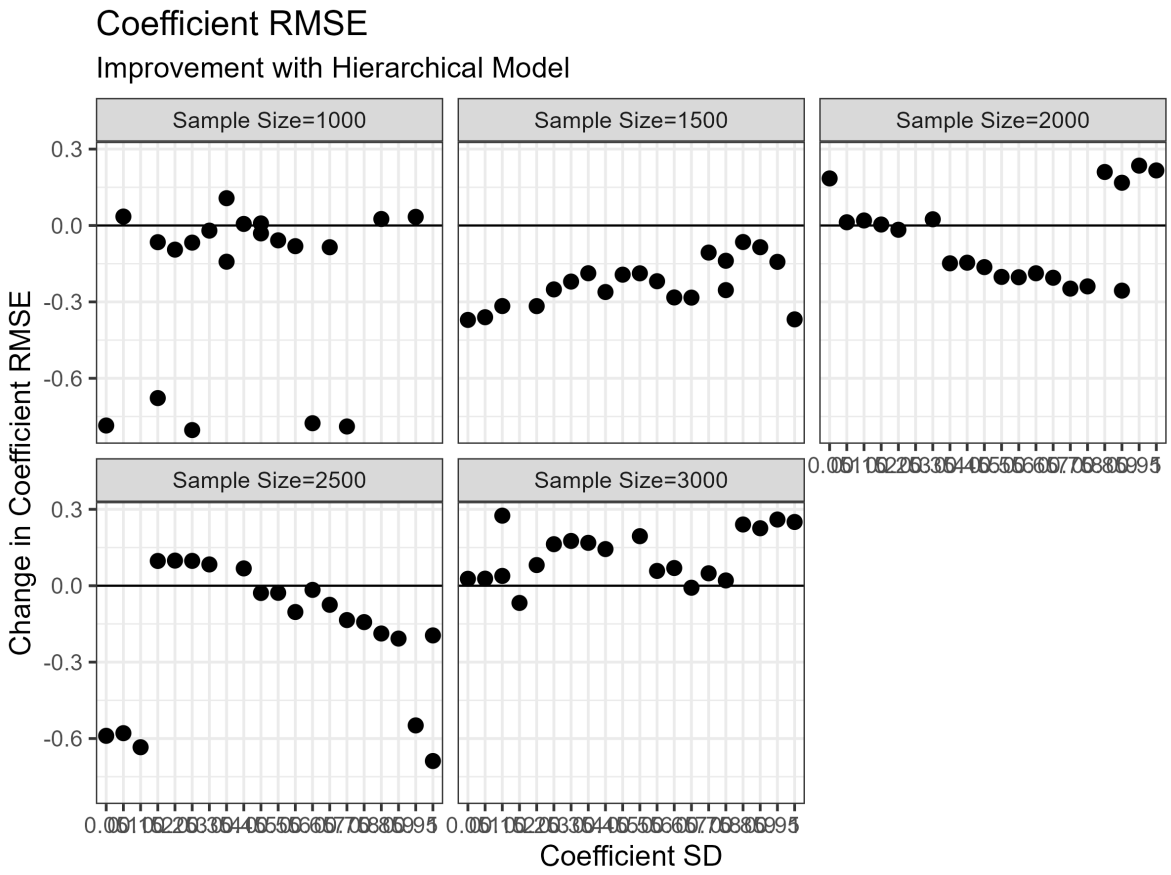

Figure 1. *Difference in root mean squared error between hierarchical and OLS estimators of heterogeneous effects with different levels of random variation in the heterogeneous effects and outcome. Positive values favor OLS, and negative values favor hierarchical modeling. Simulations based on 1,000 observations with 32 unique groups.*

values mean that the hierarchical models performs better. Because the number of observations is fixed, as the number of grouping variables rises, the size of the groups falls. For instance, with two grouping variables, there are 250 observations per group. With seven grouping variables, there are 7 or 8 observations per group.

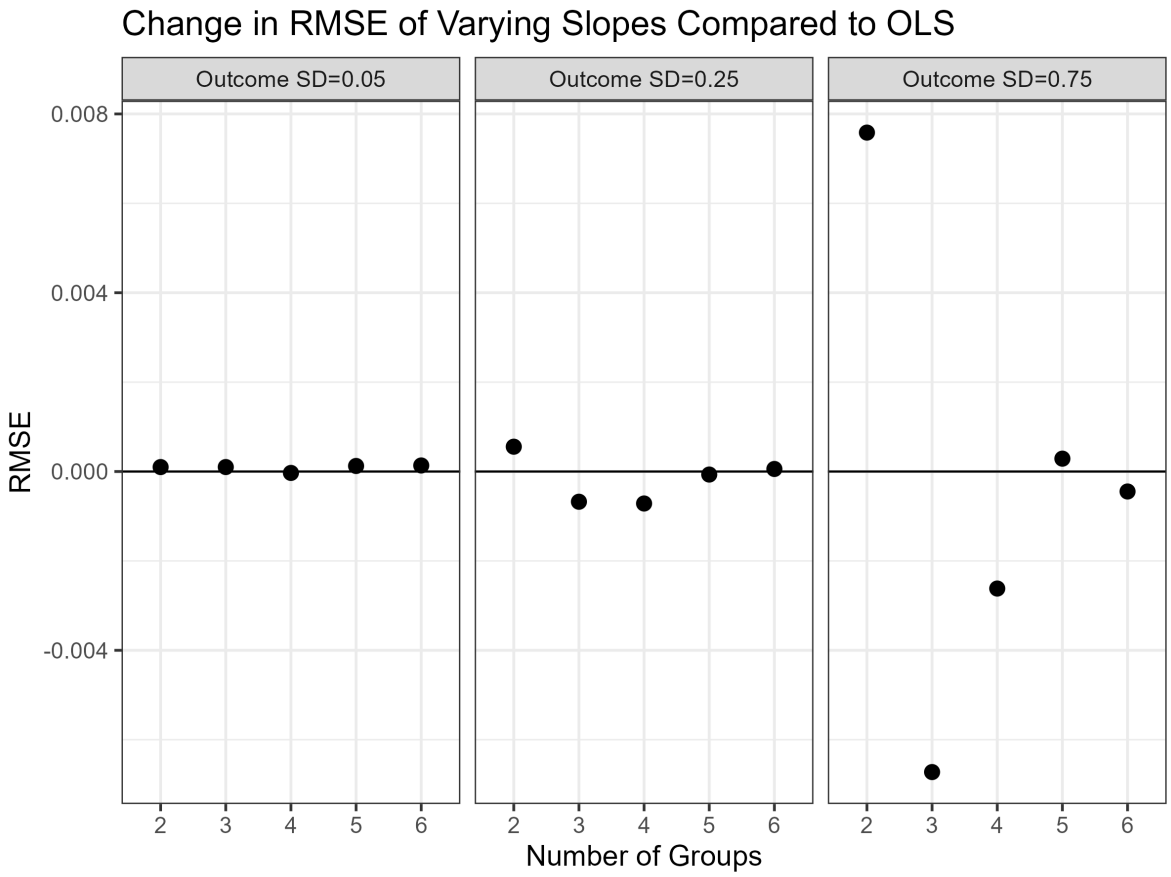

Figure 2. *Difference in root mean squared error between hierarchical and OLS estimators of heterogeneous effects with different levels of random variation in the heterogeneous effects and outcome as well as different numbers of grouping variables. Positive values favor OLS, and negative values favor hierarchical modeling. Simulations based on 1,000 observations.*

5 Example Application: Alliances and Public Support for War

In the following, I further demonstrate how the hierarchical approach works and the benefits of regularizing effect estimates by reanalyzing a study by Tomz and Weeks (2021). Tomz and Weeks (TW hereafter) examine whether the public is more willing to go to war for an allied country. In a factorial experiment with vignettes, they find a 33% average increase in support for military intervention on behalf of another country if that country is an ally. This is a large and potentially important relationship, because the United States has a global network of allies.

Given the size of the main effect, TW's paper is an ideal scenario for comparing interactions and hierarchical models. Corresponding interaction effects may be large, and their sample size of 1,200 respondents is not unusual in published work. At the same time, TW estimated an array of interactions to check how other treatments modify the impact of alliances. There are 64 unique treatment groups with anywhere from 11 to 32 respondents, so estimates of the impact of alliances in the 32 pairs of alliance treatment and control groups employ at most 54 data points. As such, employing varying slopes for regularization will likely offer substantial benefits, because the small groups will likely lead to noisy estimates. I document these gains by analyzing the how other experimental treatments modify the impact of alliances, and then exploring how demographic differences modify the alliance treatment.

5.1 Differences by Experimental Scenario

Along with alliances, TW randomly assign whether the potential beneficiary of U.S. intervention is a democracy or not, the stakes of intervention, the potential costs, and the region of the world. They estimate the impact of alliances in the 32 treatment conditions with an OLS

model that fully crosses interactions between the treatments, and calculate marginal effects that average over these groups. I keep the same fully crossed structure in the treatment interactions to define groups, but use a varying slopes model to estimate the impact of alliances in each treatment group.

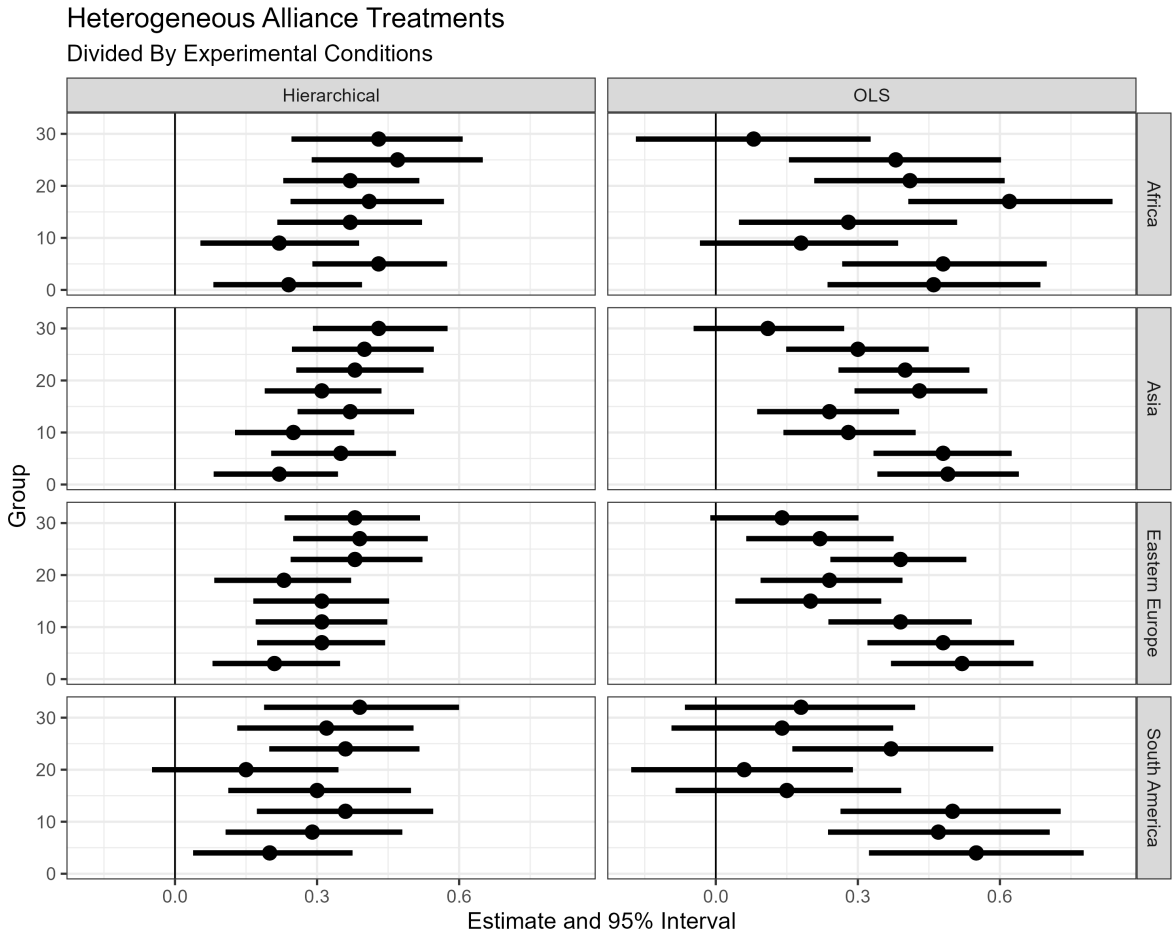

Figure 3. *Comparison of OLS and hierarchical estimates of the impact of alliance across experimental conditions. Estimates divided based on the regional treatment variable for ease of presentation. The y-axis is a numeric indicator of the groups, which there are 32 treatment-control comparisons.*

Figure 3 compares the estimated alliance treatment effects across the experimental groups with the OLS and hierarchical models. Two aspects of Figure 3 show the regularization benefits of hierarchical modeling. First, the hierarchical estimates are more precise. The credible intervals in the hierarchical model are smaller because they incorporate information from every

group.

Second, the hierarchical estimates are less variable, again due to partial pooling of the slopes. This reduces the estimated variation in how alliances impact mass attitudes, and is obvious if we compare estimates within regions. The OLS interactions are more dispersed, while the hierarchical estimates hew more closely to the overall mean. This is especially notable in the African and Latin American scenarios. Inasmuch as differences across scenarios are driven by noise in small treatment groups, the hierarchical model smooths out some of that random variation. In some cases, this leads to large shifts in the coefficient estimate.

5.2 Who Responds to Alliances

To further explore the potential application of hierarchical models, this section examines how demographic factors modify the impact of alliances. I used race, gender, hawkishness and internationalism to define demographic groups with fully crossed interactions that produce X number of groups. I selected these variables because foreign policy dispositions like militant assertiveness shape willingness to use force (Kertzer et al., 2014) as do gender (Barnhart et al., 2020) and race. I also control for other experimental manipulations.[11] Following TW's OLS analysis, I use a Gaussian likelihood, although the outcome is a binary variable.

I describe the results in three steps. First, I summarize the distribution of alliance effects in Figure 4. After this, I summarize the sources of variation in the alliance effect. Finally, I compare the hierarchical model with an equivalent OLS specification.

How alliances impact support for using force varies widely across demographic groups. Figure 4 provides an initial summary of that variation, and highlights several noteworthy estimates. Key estimates include the median, maximum and minimum effect estimates, as well as the standard deviation of all posterior draws.

First, Figure 4 notes that the minimum estimated impact of an alliance on a demographic

[11] See the appendix for priors.

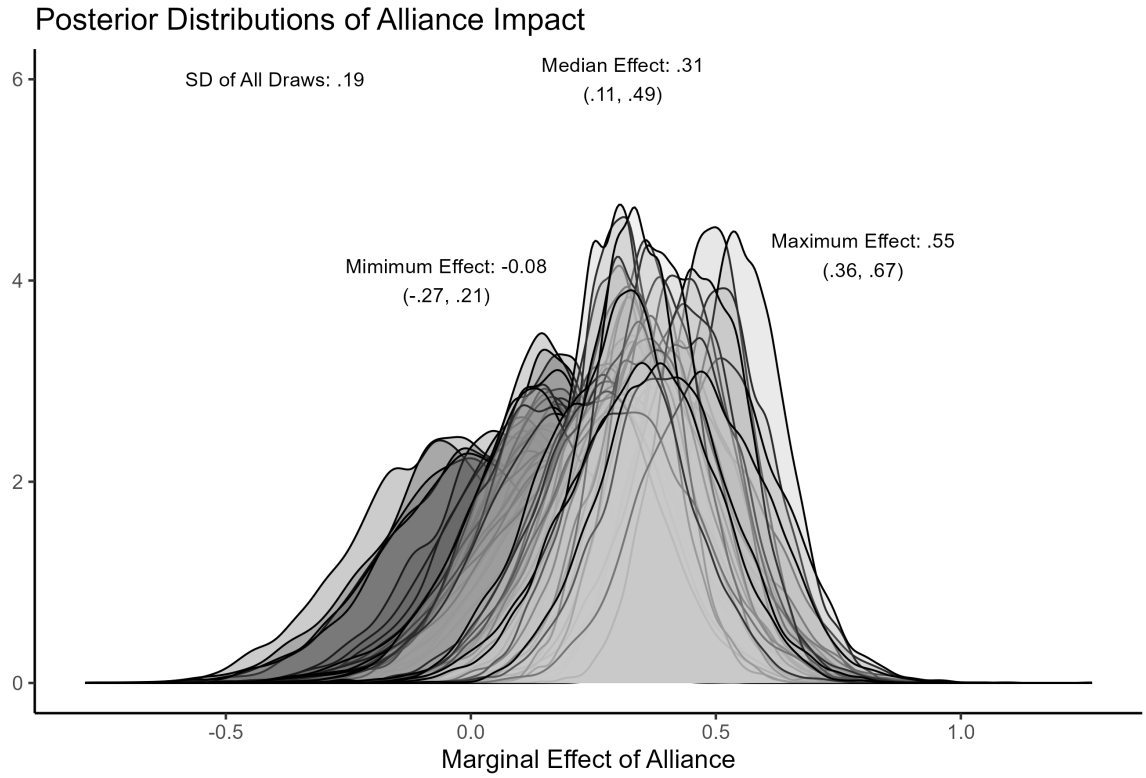

Figure 4. *Posterior distribution of all estimated impacts of alliances on support for using force. Text values give notable point estimates, and parentheses summarize the 95% credible interval.*

group is -0.08, while the maximum is .53. The maximum effect occurs among white men with high internationalism and low hawkishness. The minimum effect applies to non-white women with low internationalism and high hawkishness. There is no overlap in the posteriors of these estimates.

The median group treatment effect estimate is .31, and this group of respondents is non-white men with middling internationalism and hawkishness. This estimate is quite similar to TW's average treatment effect of .33. That average summarizes enormous demographic differences, however. The standard deviation of all posterior draws is .19.

This variation is the result of demographic differences between groups. Figure 5 plots how the impact of alliances varies across support for international engagement, willingness to use force, race and gender. Because there are many groups, the impact of alliances varies widely within each level of these variables, but there are some clear patterns.

Individuals with minimal interest in international engagement are less responsive to alliances, while any greater support for internationalism leads to a fairly consistent response to alliances. Similarly, alliances exert less impact on individuals who have minimal militant assertiveness. Alliances are very influential for individuals with low but greater than minimal hawkishness, however. Among those with moderate or high hawkishness, alliances have a fairly consistent impact. The media alliance impact is also greater for men, and among white respondents. Letting slopes vary across each level of the grouping variables generates more flexibility, and clearly shows the difference between individuals with very low internationalism or low militant assertiveness and others.

As Figure 5 suggests, alliances increase support for foreign intervention most among white men, especially those with low hawkishness and some internationalism. By contrast, alliances have little impact on support for war among non-white females who are also skeptical of international engagement and unwilling to use force. Individuals with more ambivalent foreign policy views respond more typically to TW's alliance treatment.

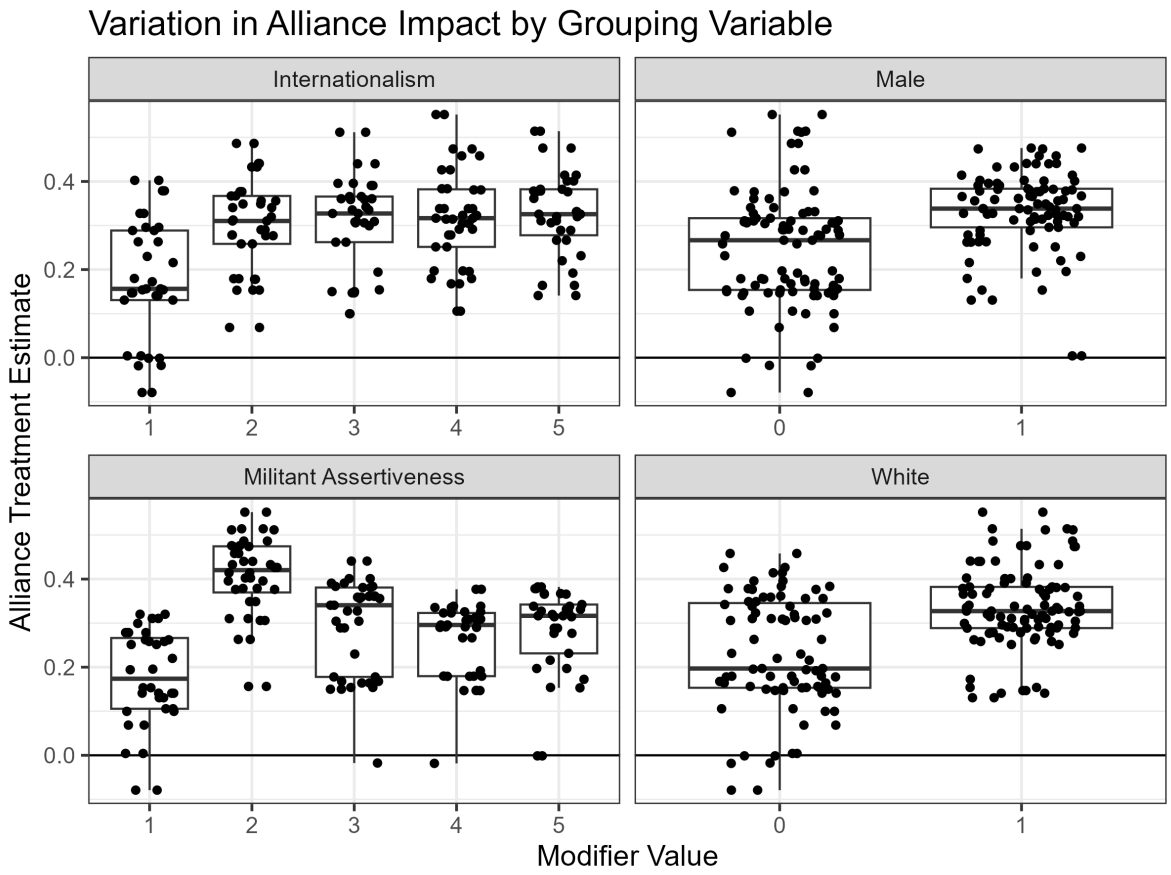

Figure 5. *Variation in the impact of alliances on support for military intervention across four variables that set groups. Each point marks the impact of alliances on a specific group, and boxplots summarize the median and interquartile range of the slopes within each level of the variable. All slopes are present in each facet.*

All these estimates suggest that internationalism matters more than than hawkishness for understanding who is willing to fight for U.S. allies. Alliances may impact hawks less because these individuals support intervention regardless. Military alliances matter most to backers of international engagement who less willing to use force, but not entirely averse to military intervention.

Finally, I illustrate the regularization benefits of a hierarchical model in Figure 6. Again, the hierarchical estimates are more precise and less variable than OLS with fully crossed interactions. This occurs because the model partially pools information across groups, which reduces uncertainty and pulls the estimated alliance impact towards the overall mean.

The other noteworthy concern is that OLS estimation with fully crossed interactions can make strong extrapolations. In some groups, the estimated impact of an alliance rises to 80 or 90%. This is the result of the purely linear estimation in OLS that stacks additive terms. Hierarchical models somewhat relax this, and are less likely to pick up noise in the data.

These results show some of the strengths and weaknesses of the hierarchical approach to heterogeneous effects.[12] A simple model based on demographic groups provides precise insights about who heeds alliances in supporting using force abroad. At the same time, because some demographic groups are small and the model pulls groups towards the overall mean, powerful comparisons between most groups is challenging. Fewer groups would have more data and less uncertainty but perhaps obscure variation across key demographic characteristics.

6 Conclusion

This note explained how and when to use hierarchical models to estimate heterogeneous effects. Bayesian modeling can apply to a wide range of outcomes, data structures, and theories. It also details what drives variation in an effect and how much an effect varies. Explicitly

[12] In the appendix, I analyze Bush and Prather (2020).

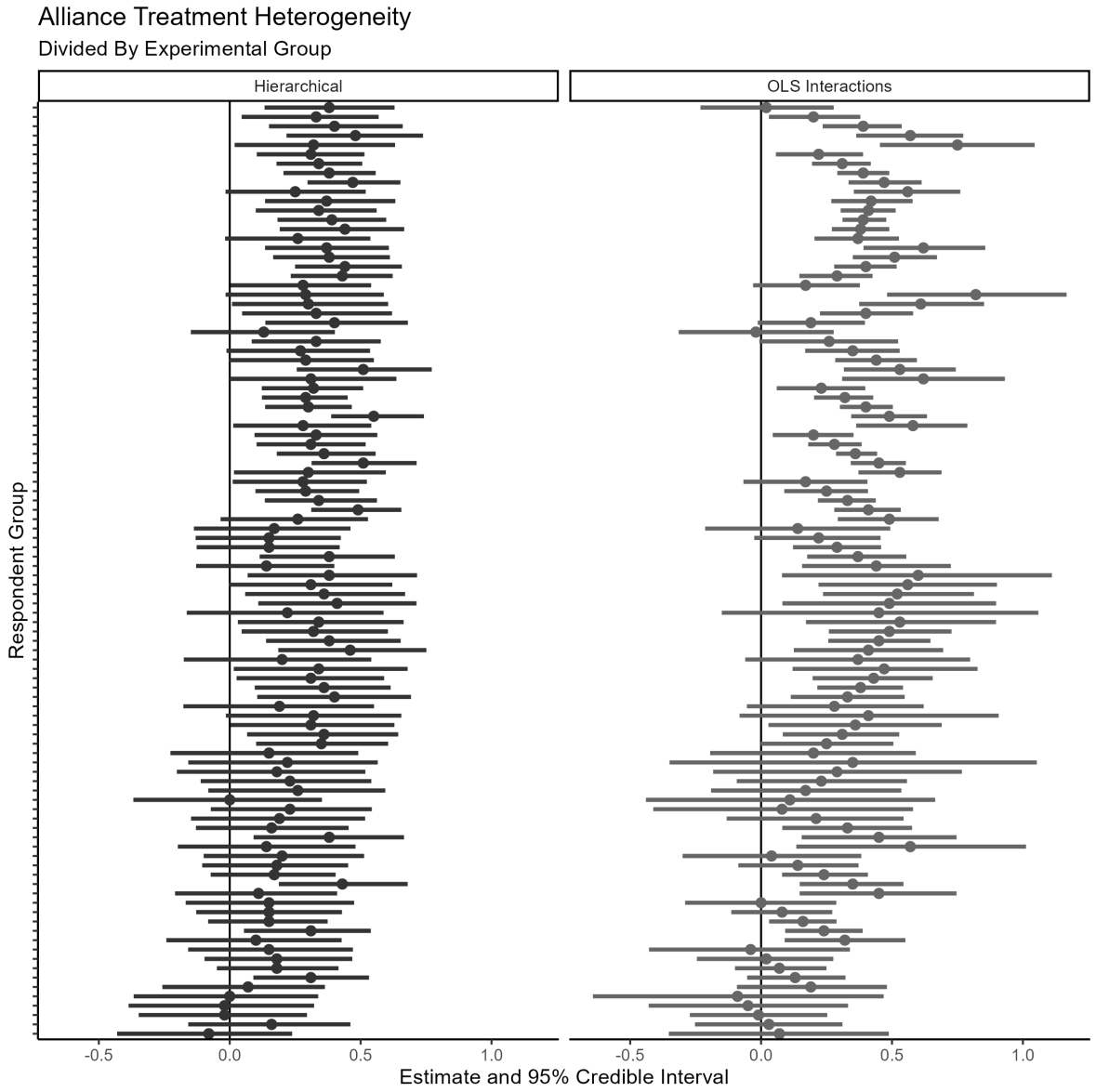

Figure 6. *Estimated impact of alliance on support for military intervention in different subgroups of respondents. Groups defined based on fully crossed interactions between the demographic variables. Error bars give the 95% credible interval in the hierarchical model and 95% confidence interval in the OLS interactions model.*

modeling how different groups respond to an independent variable can help test arguments and inform policy.

Hierarchical modeling provides an intermediate approach between interactions or subgroup analyses and machine learning algorithms. For interactions with one or two variables, relying on simple interaction tools is best. Similarly, machine learning is best for discovery of complex heterogeneity. When there are two or more modifiers and many groups of theoretical interest, hierarchical modeling allows theoretically informed and interpretable estimation of effect variation.

As a result, hierarchical modeling complements existing tools and should not replace them. Researchers can use hierarchical models to check and inform other techniques, for instance by seeing if a key interaction holds when there are multiple modifiers, or comparing multiple modifiers that past theories have identified. Using hierarchical modeling can thus help scholars and policymakers better understand heterogeneous effects.

Acknowledgements

Thanks to Taylor Kinsley Chewning, Andrew Gelman and Carlisle Rainey for helpful comments.

References

Abramson, Scott F, Korhan Koçak and Asya Magazinnik. 2022. "What Do We Learn about Voter Preferences from Conjoint Experiments?" *American Journal of Political Science* 66(4):1008–1020.

Alley, Joshua. 2021. "Alliance Participation, Treaty Depth and Military Spending." *International Studies Quarterly* 65(4):929–943.

Arel-Bundock, Vincent, Ryan C Briggs, Hristos Doucouliagos, Marco Mendoza Aviña and

Tom D Stanley. 2022. "Quantitative Political Science Research Is Greatly Underpowered." *OSF Preprints* . Available at: https://osf.io/preprints/osf/7vy2f.

Barnhart, Joslyn N, Robert F Trager, Elizabeth N Saunders and Allan Dafoe. 2020. "The Suffragist Peace." *International Organization* 74(4):633–670.

Blackwell, Matthew and Michael P Olson. 2022. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30(4):495–514.

Bürkner, Paul-Christian. 2017. "brms: An R package for Bayesian multilevel models using Stan." *Journal of Statistical Software* 80(1):1–28.

Bush, Sarah Sunn and Lauren Prather. 2020. "Foreign Meddling and Mass Attitudes Toward International Economic Engagement." *International Organization* 74(2):584–609.

Chaudoin, Stephen. 2014. "Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions." *International Organization* 68(1):235–256.

Clark, William Roberts and Matt Golder. 2023. *Interaction Models: Specification and Interpretation*. Cambridge University Press.

Clifford, Scott and Carlisle Rainey. 2023. Estimators for Topic-Sampling Designs. Technical report.

Clifford, Scott and Carlisle Rainey. 2024. "Estimators for Topic-Sampling Designs." *Political Analysis* p. 1–14.

Dorie, Vincent, George Perrett, Jennifer L Hill and Benjamin Goodrich. 2022. "Stan and BART for Causal Inference: Estimating Heterogeneous Treatment Effects Using the Power of Stan and the Flexibility of Machine Learning." *Entropy* 24(12):1782.

Feller, Avi and Andrew Gelman. 2015. "Hierarchical Models for Causal Effects." *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource* pp. 1–16.

Gelman, Andrew. 2008. "Scaling regression inputs by dividing by two standard deviations." *Statistics in medicine* 27(15):2865–2873.

Gelman, Andrew. 2018. "You need 16 times the sample size to estimate an interaction than to estimate a main effect.". Available at: https://statmodeling.stat.columbia.edu/2018/03/15/need16/.

Goplerud, Max. 2021. "Modelling Heterogeneity Using Bayesian Structured Sparsity." *arXiv preprint arXiv:2103.15919* .

Green, Donald P and Holger L Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413–434.

Guisinger, Alexandra and Elizabeth N. Saunders. 2017. "Mapping the Boundaries of Elite Cues: How Elites Shape Mass Opinion across International Issues." *International Studies Quarterly* 61(2):425–441.

Hopkins, Vincent, Ali Kagalwala, Andrew Q Philips, Mark Pickup and Guy D Whitten. 2024. "How Do We Know What We Know? Learning from Monte Carlo Simulations." *The Journal of Politics* 86(1):36–53.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Kertzer, Joshua D., Kathleen E. Powers, Brian C. Rathbun and Ravi Iyer. 2014. "Moral Support: How Moral Values Shape Foreign Policy Attitudes." *The Journal of Politics* 76(3):825–840.

Kertzer, Joshua D and Ryan Brutger. 2016. "Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory." *American Journal of Political Science* 60(1):234–249.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the national academy of sciences* 116(10):4156–4165.

Levendusky, Matthew S and Michael C Horowitz. 2012. "When Backing Down is the Right Decision: Partisanship, New Information, and Audience Costs." *The Journal of Politics* 74(2):323–338.

Marquardt, Kyle L. 2022. "Language, Ethnicity, and Separatism: Survey Results from Two Post-Soviet Regions." *British Journal of Political Science* 52(4):1831–1851.

McElreath, Richard. 2016. *Statistical Rethinking: A Bayesian course with examples in R and Stan*. CRC Press.

Schwartz, Joshua A and Christopher W Blair. 2020. "Do Women Make More Credible Threats? Gender Stereotypes, Audience Costs, and Crisis Bargaining." *International Organization* 74(4):872–895.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22(11):1359–1366.

Tomz, Michael and Jessica L.P. Weeks. 2021. "Military Alliances and Public Support for War." *International Studies Quarterly* 65(3):811–824.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.