Using Hierarchical Models to Estimate Heterogeneous Effects

Joshua Alley
Assistant Professor
Baylor University
Joshua_Alley@baylor.edu

February 6, 2026

Abstract

This paper 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, and starts with identify the key sources of heterogeneity. Then, researchers should 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 and partially pools group estimates. This captures systematic and random 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. The downside is that this approach employs very strong regularization, which breaks down if there are many small groups. I document these claims with a simulation analysis and extension of a published study.

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 an 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 distinct modifiers for the same independent variable 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.

[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.

Such proliferation 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 the most common tool. Simple interactions and subgroup analyses are ubiquitous because they are relatively easy to interpret, but they have serious power concerns. Many political science analyses have low power even to detect main effects (Arel-Bundock et al., 2025). 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).

Even when theory implies that scaling up the number of modifiers is necessary, doing so with interactions is not easy. The interpretation benefits of interactions diminish as researchers add modifiers. Adding variables to an interaction 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, however, especially in smaller social science datasets. A lack of interpretability is especially problematic for testing the relative weight of multiple modifiers.

The hierarchical strategy I propose here addresses the power shortcomings of interactions

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

while retaining more theoretical structure than machine learning. I do this by showing how scholars can use two connected regressions to estimate theoretically informed models of 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 and regularization will break down if there are many small groups. It also may not scale to very large datasets, depending on the underlying sampler. Hierarchical modeling therefore works best when theory indicates more than two modifying factors and there is less emphasis on discovery.[3]

Hierarchical modeling has an additional benefit of bridging the disparate estimates of interactions and machine learning. Many interactions functionally estimate grouped effects—individuals with the same values of the modifiers will have the same effect estimate. Machine learning models usually estimate individual effects. The model I propose here gives individual estimates with the systematic group component and a way to account for and regularize individual deviations from those groups.

There are two key steps when theory and data make using hierarchical models worthwhile. First, researchers should identify potential modifiers of a treatment and use them to model treatment effects. Second, they should take that model of treatment effects and connect it to a model linking individual treatment effects and the 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. These are crucial advantages in a world with many conditional theories.

[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.

While frequentist estimation of hierarchical models is possible, Bayesian estimation is easy, usually fast, 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 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 exert the strongest impact on respondents with high internationalism and interest in the news. It also documents the importance of regularization in analyzing subgroups derived from combinations of experimental treatments.

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 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

[4] I provide example code below and in the appendix.

[5] It also facilitates pre-registration when applicable.

sources of heterogeneity will obfuscate results and can hinder model fitting.

There are three general approaches to defining key modifiers. First, researchers can use 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 could include indicators of issues, for instance. A similar application of hierarchical estimators for topic-sampling experiments estimates how a treatment effect varies across different topics (Clifford and Rainey, 2023).

The most common practice in estimating heterogeneous treatment effects is fully crossed interactions. This estimates the impact of a treatment across experimental strata, but risks spurious results by functionally estimating subgroup results. Most social science experiments do not have adequate power for main effects (?), let alone small subgroups that may only include 50 or fewer data points.

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. Researchers could use a mix of individual and contextual factors to predict divergent consequences of a survey experiment treatment. Such a model might include factors such as an indicator of state of residence, age, gender, and race.[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 (Martin, 2005), so this model of heterogeneity accounts for potential confounding, and finds

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

that after accounting for depth, democracy does not impact the relationship between alliances and defense spending, contrary to DiGiuseppe and Poast (2018).

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 race, gender and region, plus additional controls or other factors.

After defining moderators and how they relate to individual effects, 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.

[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.

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

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 an outcome; linear combinations of variables, interactions, or other terms such as varying intercepts. 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 $\mathbf{Z}$. Individual 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

[9] In brms, using non-linear syntax can express a model with a treatment, two controls, and three modifiers as: y ~ lambda * treat + controls, lambda ~ mod1 + mod2 + mod3, controls ~ coutrol1 + control2, nl = TRUE

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.

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. Extremely complex specifications can lead to model fitting problems. Sometimes, fitting problems indicate that the model is misspecified, so these problems can be a blessing in disguise.

Furthermore, hierarchical models can show general trends, but will not make powerful comparisons between every treatment effect. Researchers who want to compare specific effects

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

will often lack empirical leverage. 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. Researchers should still remember power concerns with interactions, however.

With two or more modifiers, hierarchical models begin to add value beyond. Interpreting triple interactions between a variable and two modifiers is challenging. The advantages of hierarchical modeling increase with the number of modifiers, so long as model complexity does not hinder accurate fitting and simulation from the posterior.

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.

Data size is relevant to model selection as well. Machine-learning approaches work especially well with large, often massive datasets. Interactions also benefit from more data, as it can help overcome power problems. Hierarchical models also benefit from more data, but are less sensitive to outliers in small samples. In very large datasets, some hierarchical models may take hours to fit.

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

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

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

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 this hierarchical model compares to interactions in OLS models, I first assess their performance on simulated data. Each simulation approximates the two most common applications of these models. The first simulates estimating heterogeneous treatments in factorial experiments that have an even number of data points in each group. The second simulation deals with imbalanced groups, as would likely be the case in demographic analyses.

Both simulations fix the number of observations at 2,000 and manipulate the number of groups. In each simulation, the data generating process includes group-specific effects, drawn from a normal distribution with a mean of .2 and variance of .3. I also add three control variables with fixed coefficients the predict the mean of the simulated outcome. The simulated data has a standard deviation of one.

To compare the models, I use the root-mean squared error of the coefficient estimates compared to the true value. I use root mean squared error because the hierarchical model's reduction of variance may introduce bias (Clifford and Rainey, 2024). Less variance may overcome greater bias in some coefficients and make the hierarchical model more accurate on average.

In the first simulation, I divide the 2,000 observations into equally sized combinations of three to seven binary variables, each with equal probability. This creates 8 to 128 unique groups, with 250 observations per group at the low end and 16 observations at the high end. I then fit OLS models with fully crossed interactions of the group variables, as these estimate treatment effects in each subgroup and are a common approach. In the hierarchical model, I also fully cross interactions of the group variables and set a random intercept for each group.

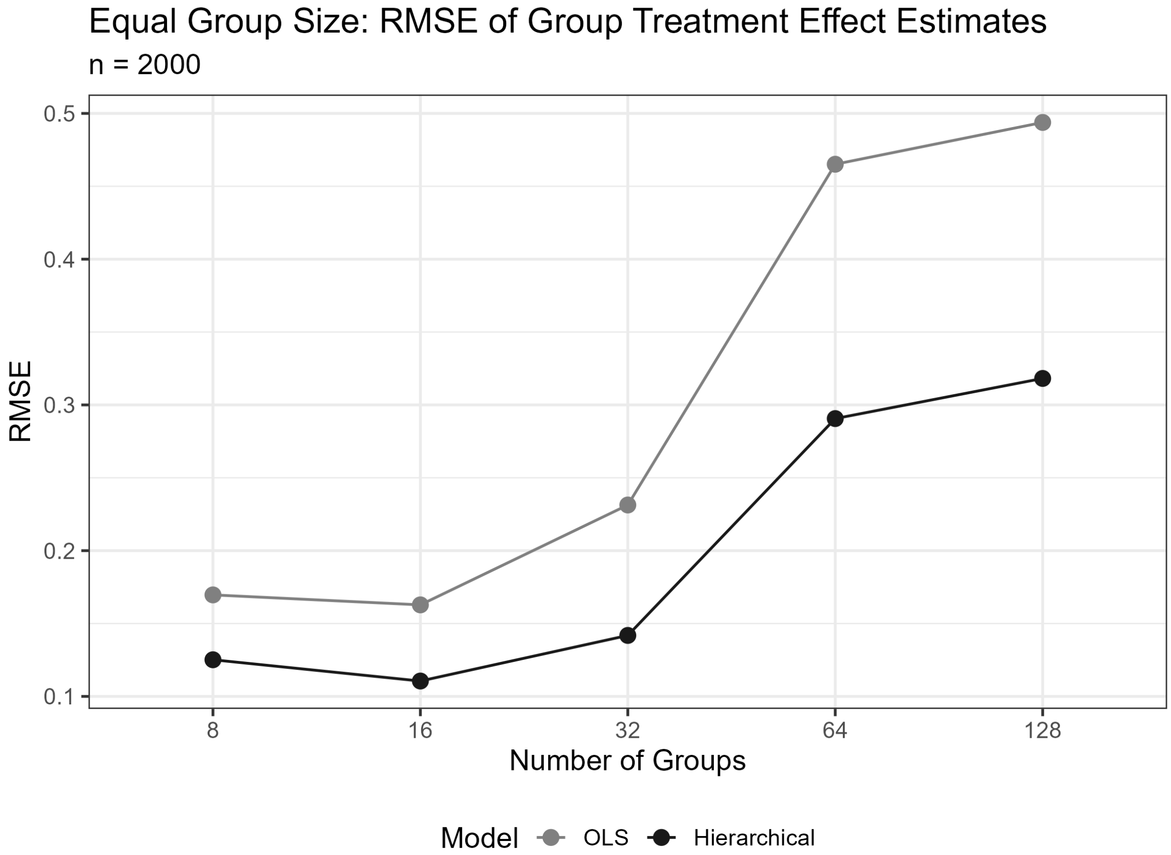

Figure 1. *Comparison of OLS and hierarchical root mean squared error in group coefficient estimates. Simulation fixes the sample size at 2,000 observations and varies the number of groups. Each group has the same number of observations.*

Figure 1 presents the results of this first simulation. When the groups are large because there are thee grouping variables, there is a small gap in model performance. The hierarchical model is marginally better. But as group numbers rise and the number of observations in each group falls, root mean squared error rises for all models, but the OLS with fully crossed interactions performs much worse. With 128 groups and 16 observations per group, the OLS model has a root mean squared error of .5, while the hierarchical model is at .33.

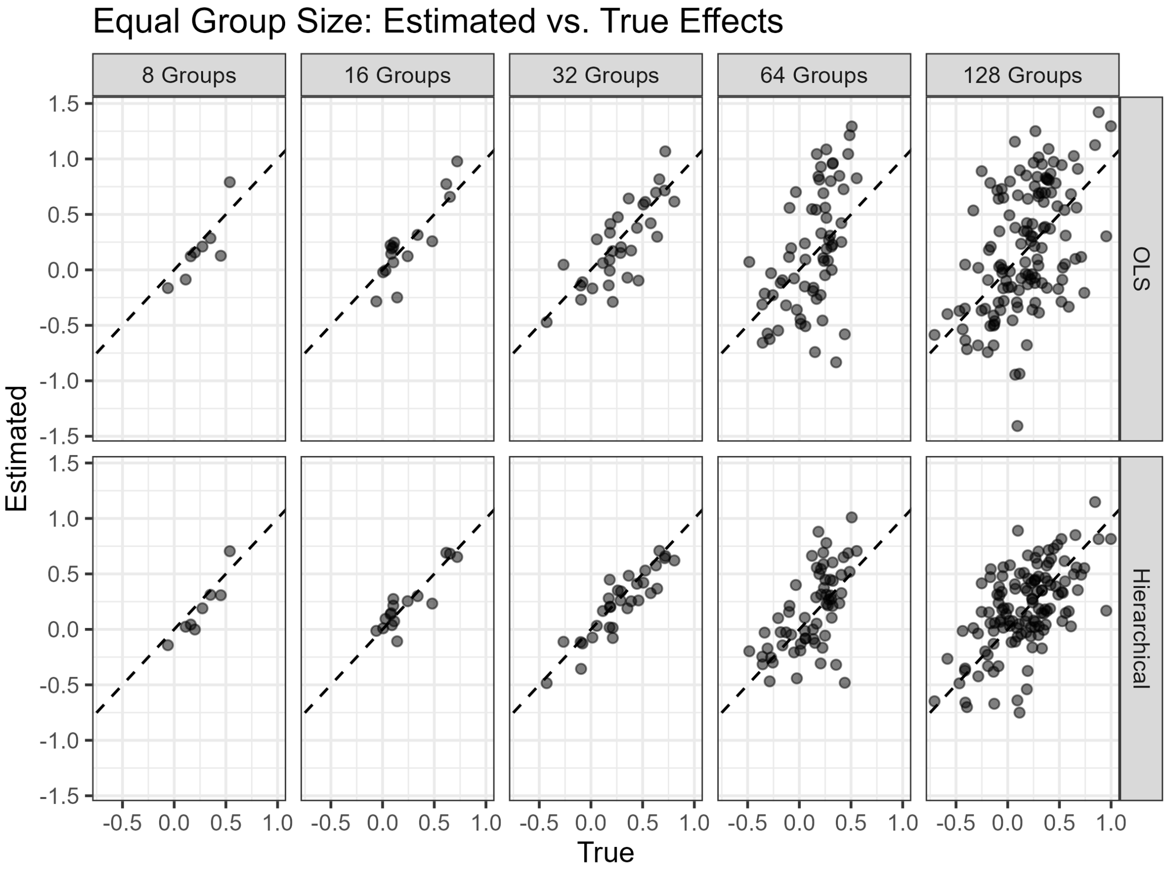

Figure 2. *Comparison of OLS and hierarchical estimates of group-specific treatment effects. Points on the dashed line are unbiased.*

These benefits are due to strongly reduced variance in the hierarchical estimates, as Figure 2 shows. This figure plots the true group treatment effects against the model estimates. Points above the dotted line are over-estimated by the model compared to the true value, while points below are under-estimated.

Even with 8 groups and 250 observations per group, more OLS estimates are unbiased, but two groups are dramatically off. This means that even with slight bias for many groups, the hierarchical estimates have lower overall root mean squared error. As the number of groups increases and group size falls, the estimates for both models become more noisy, but the hierarchical model estimates are far less variable.

This is essential because the dramatic over and under-estimates in the OLS model are the subgroup estimates with the greatest likelihood of statistical significance. But most of these estimates are very biased and reflect noise in the data, not a true effect. The problems this poses for replication are now well understood, and the hierarchical model offers a clear solution.

The second simulation shows similar improvements in an application with unbalanced group sizes. Here, I set the number of groups to 64, using six binary grouping variables. I then vary the probability of each group level and the symmetry of these probabilities across groups. Balanced groups are equivalent to the 64 group case in the first simulation. At the strongest imbalance, one group has a 20% share of ones and 80% zeros, and another has the opposite extreme, with a range in between. Such extreme imbalance can leave as few as one observation in each group, so it is an extreme test of model performance. Again, I compare the models with the root mean squared error of the coefficient estimates, and show these results in Figure 3.

Under balanced groups, the hierarchical model has a significant advantage. Greater imbalance reduces this improvement, but the hierarchical model has far lower root-mean squared error. The smaller and more numerous the groups, the less valuable the regularization becomes, however. Small groups add far less information and are strongly pulled to the overall mean, which itself is a function of the larger groups. Again, the strong imbalance creates conditions where it is wiser to set up less finely-tuned groups or expand the sample size so groups have more observations and information to contribute to partial pooling.

To further illustrate how group size shapes the hierarchical estimates, I plot the bias of the

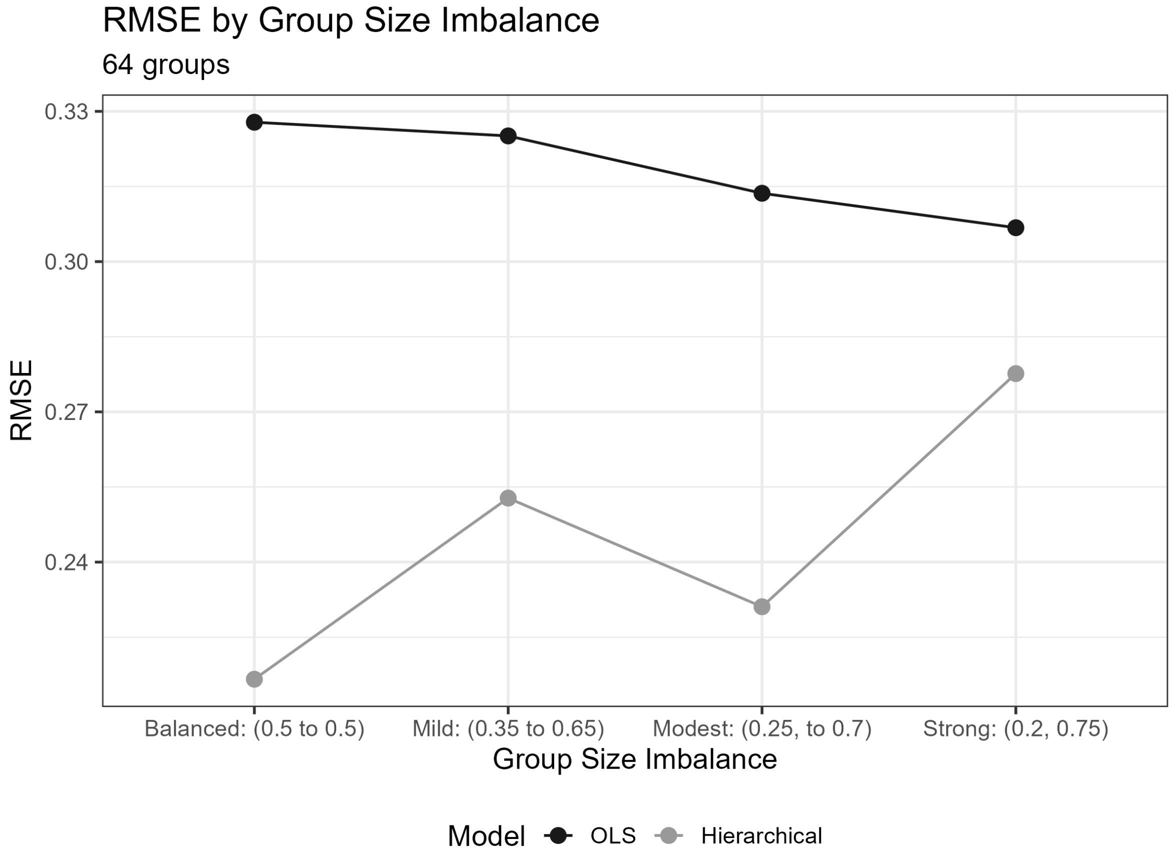

Figure 3. *Comparison of OLS and hierarchical root mean squared error in group coefficient estimates. Simulation fixes the sample size at 2,000 observations and varies the relative size of different groups.*

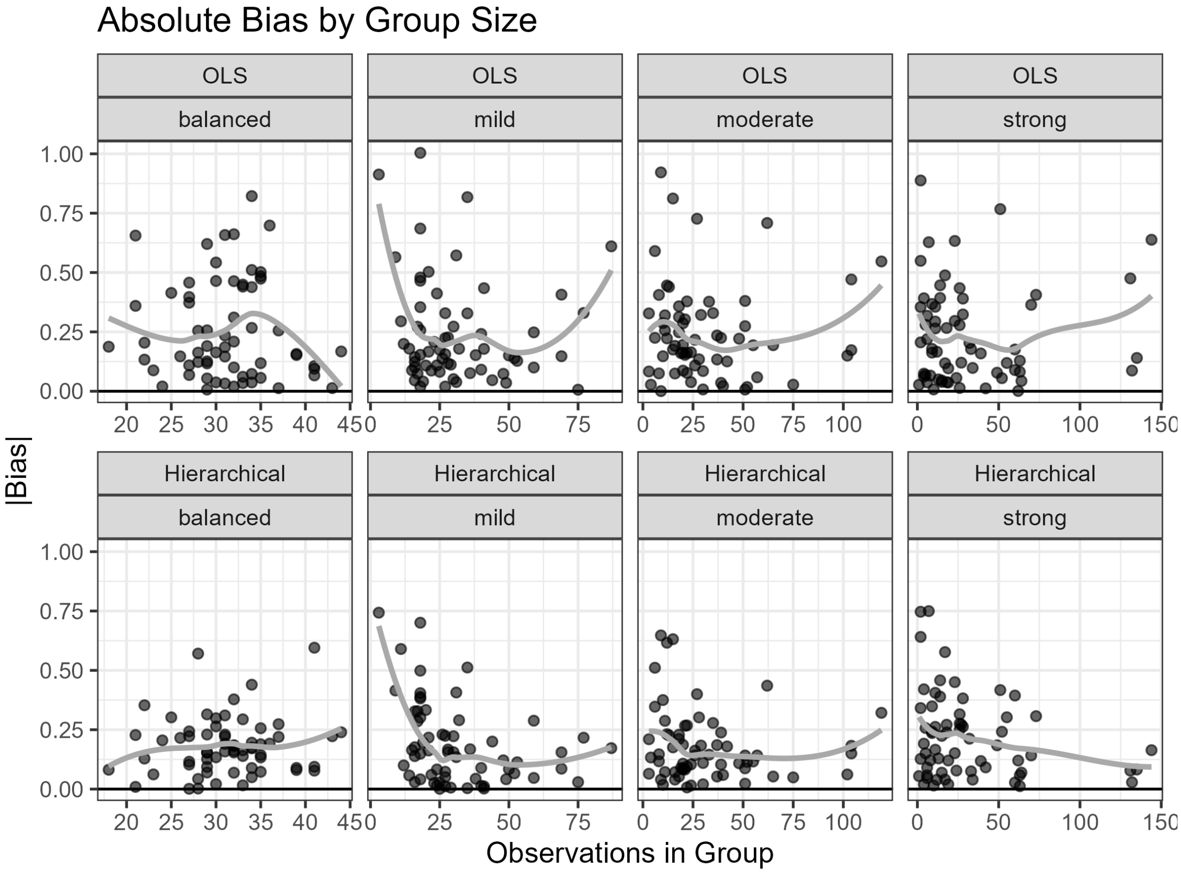

Figure 4. *Comparison of OLS and hierarchical model bias in estimates of group-specific treatment effects. Group size gives the number of observations in each group, and bias is expressed as an absolute value. Loess lines give average bias across the range of observations.*

OLS and hierarchical models in Figure 4. Again, the OLS estimator has significant bias for some groups as the model picks up noise. The magnitude of the bias does not depend on group size. In fact, large groups often have more bias because the OLS estimates weight each group equally.

The hierarchical model has much less bias for larger groups. Smaller groups with less than 20 data points have higher bias, while large groups, especially those with 75 or more observations, have much less data. This is partial pooling in action.

The simulations thus give two crucial inferences. First, partial pooling and modeling treatment effects offers significant improvements in inference, as it reduces variance in the estimates. The relative size of groups shapes the magnitude of this benefit, as hierarchical models will be less biased for large groups and more biased with small groups under extreme imbalances in group size. This means that the smaller the group, the less a hierarchical model can tell you, but it will likely be more accurate than OLS.

5 Example Application: Alliances and Public Support for War

In the following, I 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, hierarchical regularization will likely offer substantial benefits, because the small groups will likely lead to noisy estimates. I document these gains by analyzing 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 over different averages of these groups. I use a hierarchical model to estimate the impact of alliances, with fully crossed experimental treatments as the systematic modifiers of the alliance effect. This mirrors TW's model, but adds a regularization component to the heterogeneous effects of alliances.

Figure 5 compares the estimated alliance treatment effects across the experimental groups with the OLS and hierarchical models. This figure illustrates the regularization benefits of hierarchical modeling by showing the difference between the median hierarchical estimate and the OLS estimate for each combination of treatment groups and plotting the distribution of effects. Both facets give a sense of when and how regularization is useful.

First, the hierarchical estimates are much less variable, despite producing many more estimates. The hierarchical model estimates 602 non-zero alliance treatment effects, while the OLS model has 184 unique estimates, yet the hierarchical estimates are much more tightly clustered around the overall mean. This occurs because the OLS model mechanically forces individual

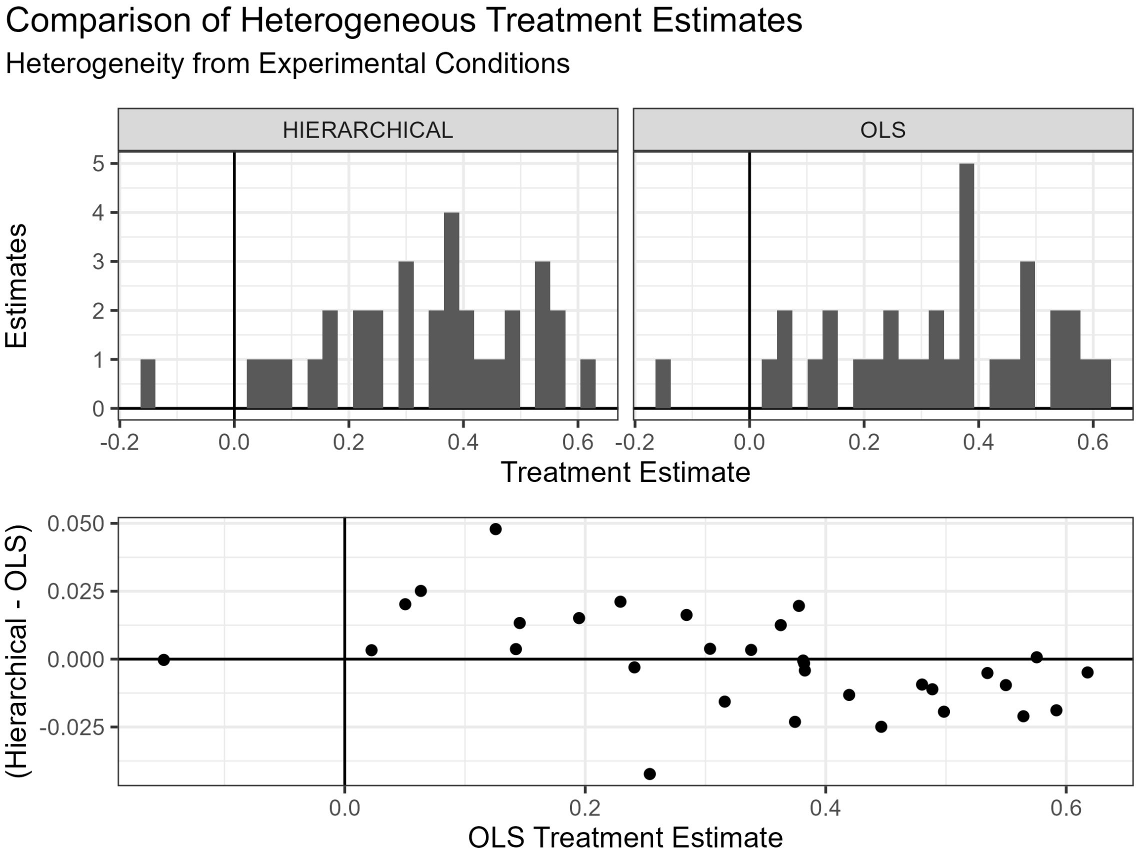

Figure 5. *Comparison of OLS and hierarchical estimates of the impact of alliance across experimental conditions. The top panel gives a histogram of the treatment effects from each model. The bottom panel gives the difference between the hierarchical and OLS estimates for each group.*

treatment effect estimates to the value of the coefficients and corresponding subgroup means point. Given the size of this sample, the treatment effects are based on comparisons of roughly 25 treatment and 25 control respondents.

Second, the hierarchical estimates pull in unusual values. The bottom panel of Figure 5 shows the difference between the median hierarchical estimate among individuals and a group and the OLS estimate for that group, which applies to all individuals in each group. The downward slope from positive to negative values means that the hierarchical estimates are larger than below average OLS estimates, but smaller than above average interaction estimates. This occurs because hierarchical estimates are pulled towards the overall mean, away from extreme and implausible values such as a 60% effect for alliances when there are low stakes, low costs and a democratic partner in Africa.

Partially because it regularizes the estimates, the hierarchical model also leads to slightly different conclusions about when alliances matter most. I compare the estimates for each group more precisely in Figure 6. For some conditions, such as a high-stakes, high-cost intervention to support a democracy, the hierarchical estimate is as much as 30% greater than the OLS estimate.

In other groups, regularization pulls down very large estimates. For instance, TW highlight a very large positive effect of alliances on support for low stakes, low cost interventions to help an autocracy, but that effect is closer to 20% in the hierarchical models, not the 46% effect that TW report.

These changes, which deal with a very large average effect, do not change the direction of the findings- in some scenarios they even strengthen them. However, the magnitudes of some subgroup estimates do matter, as they might lead observers to misidentify when alliances matter most. For other, smaller effects, regularization with a hierarchical model might change the direction of a subgroup analysis.

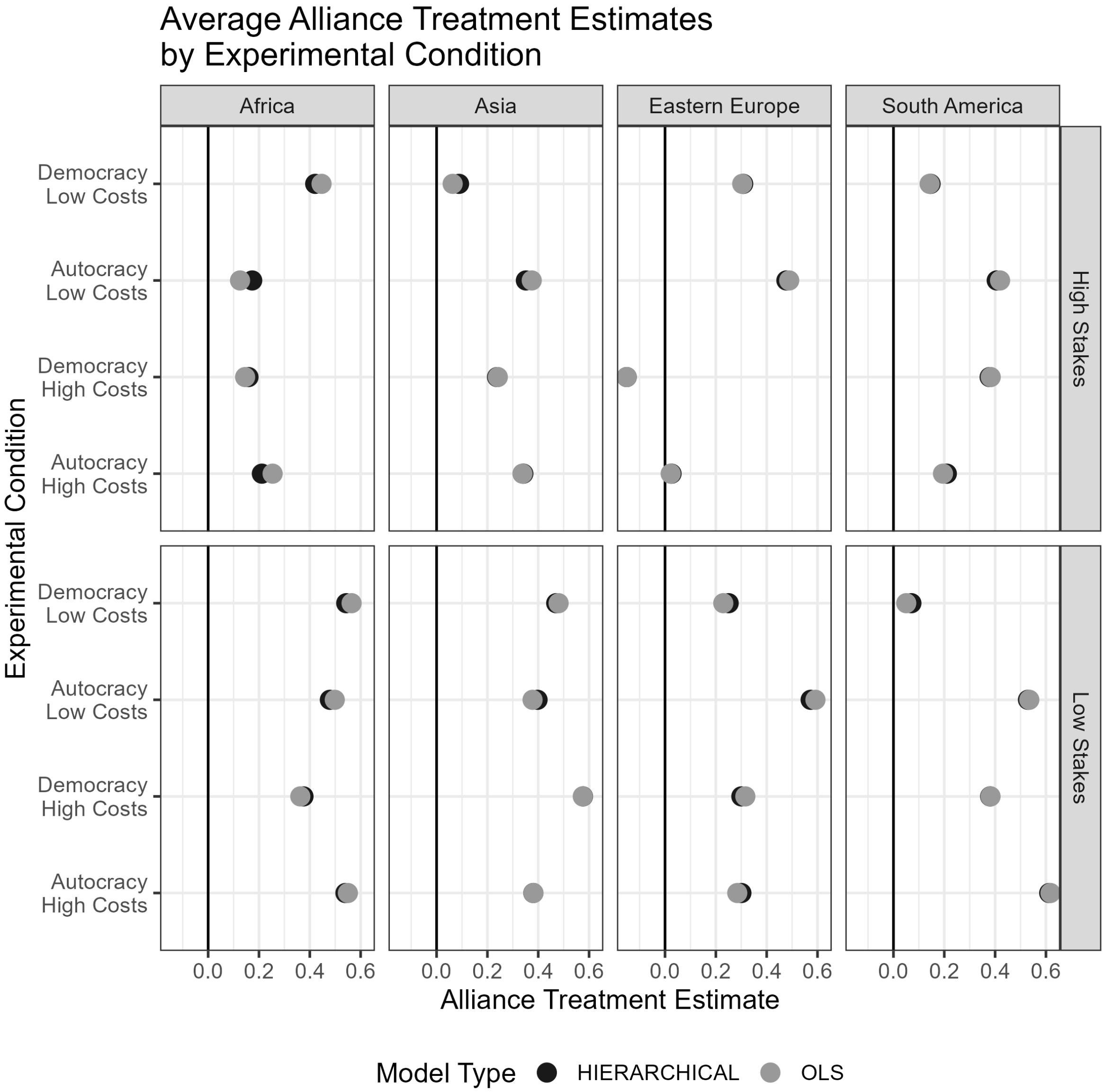

Figure 6. *Comparison of OLS and hierarchical estimates of the impact of alliance across experimental conditions. Each point gives the typical effect estimate, specifically the median of the individual effects in the hierarchical model.*

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 party, political interest, race, gender, hawkishness, and internationalism to modify the impact of alliances. 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, while Tomz and Weeks examine party and political interest as potential modifiers. I control for other experimental manipulations.

The resulting hierarchical model thus encompasses four subgroup comparisons in TW's appendix, as well as additional information about other demographic factors. Doing all of the comparisons encapsulated in the hierarchical model would require at least six pairwise interaction models. Following TW's OLS analysis, I use a Gaussian likelihood, although the outcome is a binary variable.

To start, I plot the two regressions of hierarchical model. Figure 7 plots how different variables shape either the impact of an alliance or support for intervention. One facet of this figure gives the coefficient estimates from the model of heterogeneity, and a second gives the control estimates.

Internationalism and news interest most accentuate the impact of alliances on support for war. A one-unit increase in internationalism increases the impact of the alliance treatment on support for war by .1 in expectation, with a 95% credible interval that ranges between .07 and .14. Hawkishness is also correlated with higher support for war, but not as largely or clearly. Respondents with high news interest also respond to alliances by .1 more than others in this survey, and that relationship could be as small as .03 or as large as .18. This differs from TW's conclusion that respondents with high and low news interest respond in roughly the same way, likely because the hierarchical model accounts for other factors that are correlated with news interest.

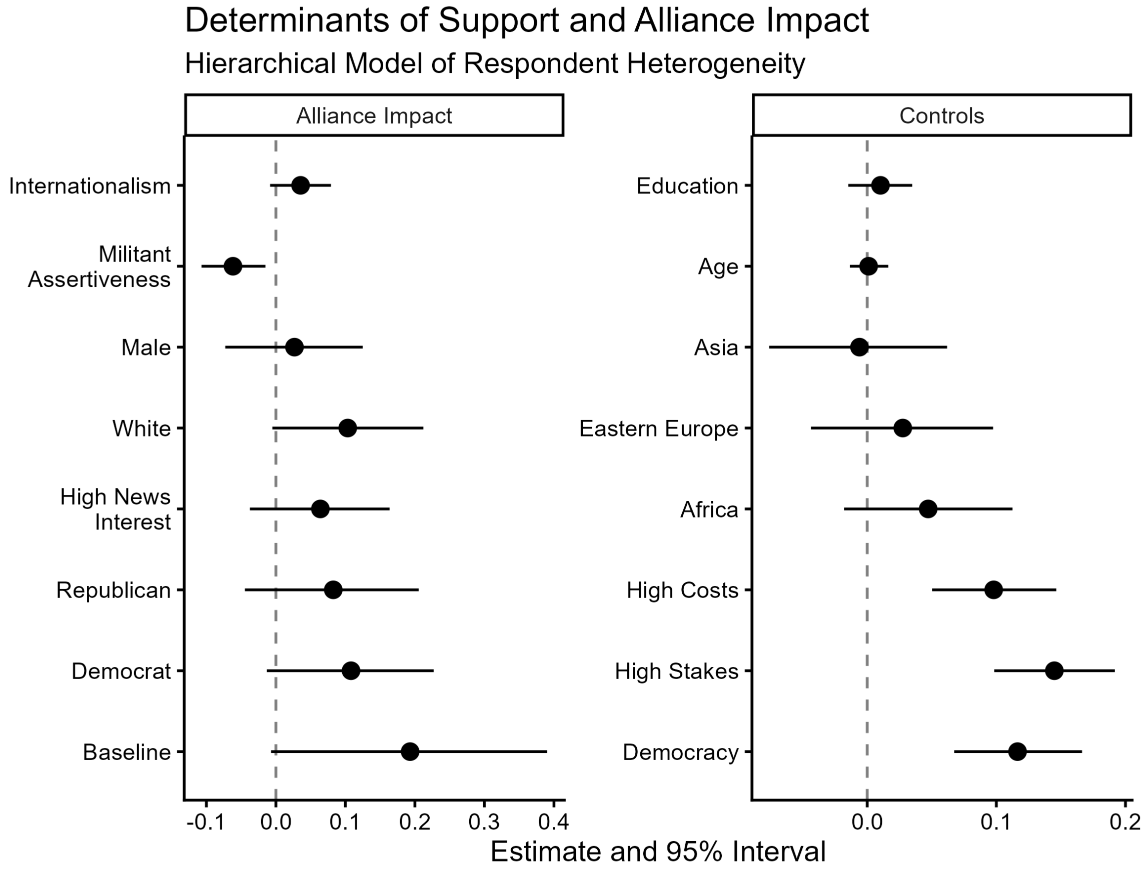

Figure 7. *Variation in the impact of alliances on support for military intervention across four variables that set groups. Each point marks the impact of alliances on a specific group, and boxplots summarize the median and interquartile range of the slopes within each level of the variable. All slopes are present in each facet.*

Other demographic predictors do not differentiate responses to alliances as clearly. The impact of alliances may be somewhat greater for white and male respondents, but those differences are not clearly different from zero as the 95% credible intervals include zero and small negative values. Similarly, Republicans and Democrats are not all that different from independents in how they respond to alliances. This matches TW's conclusion that partisanship does not create substantial differences in alliance effects. The intercept in this equation can be theoretically meaningful, but it is not in this case because neither internationalism nor militant assertiveness take a value of zero.

The control equation suggests that high stakes, high costs, and partner democracy all increase support for intervention. Region is less consequential, as neither Asia, Eastern Europe nor Africa is clearly different from the reference category of Latin America. As with any regression, the intercept here gives the impact of alliances when all the other variables are equal to zero, and this indicates that alliances increase support for war by .26 given low costs, low stakes and an autocracy in Latin America. The 95% credible interval of this intercept estimate includes the corresponding hierarchical estimate of the alliance treatment in ??.

While it is possible to construct a rough profile of who responds most to alliances using the estimates in Figure 7, I provide a more precise summary in Figure 8. This figure plots the median estimate in each group of respondents, after grouping respondents based on race, gender, news interest and foreign policy dispositions. Darker fill indicates a positive effect that has a credible interval excluding zero and negative values. Some combinations of demographic data are blank because they are missing from the data.

Alliances most impact white men with high internationalism, high militant assertiveness, and high interest in the news. The largest median estimate of an alliance impact is .7. At each level of internationalism, alliance impact increases by .1, which was the coefficient value in the heterogeneity regression. News interest generates a similar shift in the median alliance impact across all groups. The marginal role of hawkishness is also apparent in substantial differences

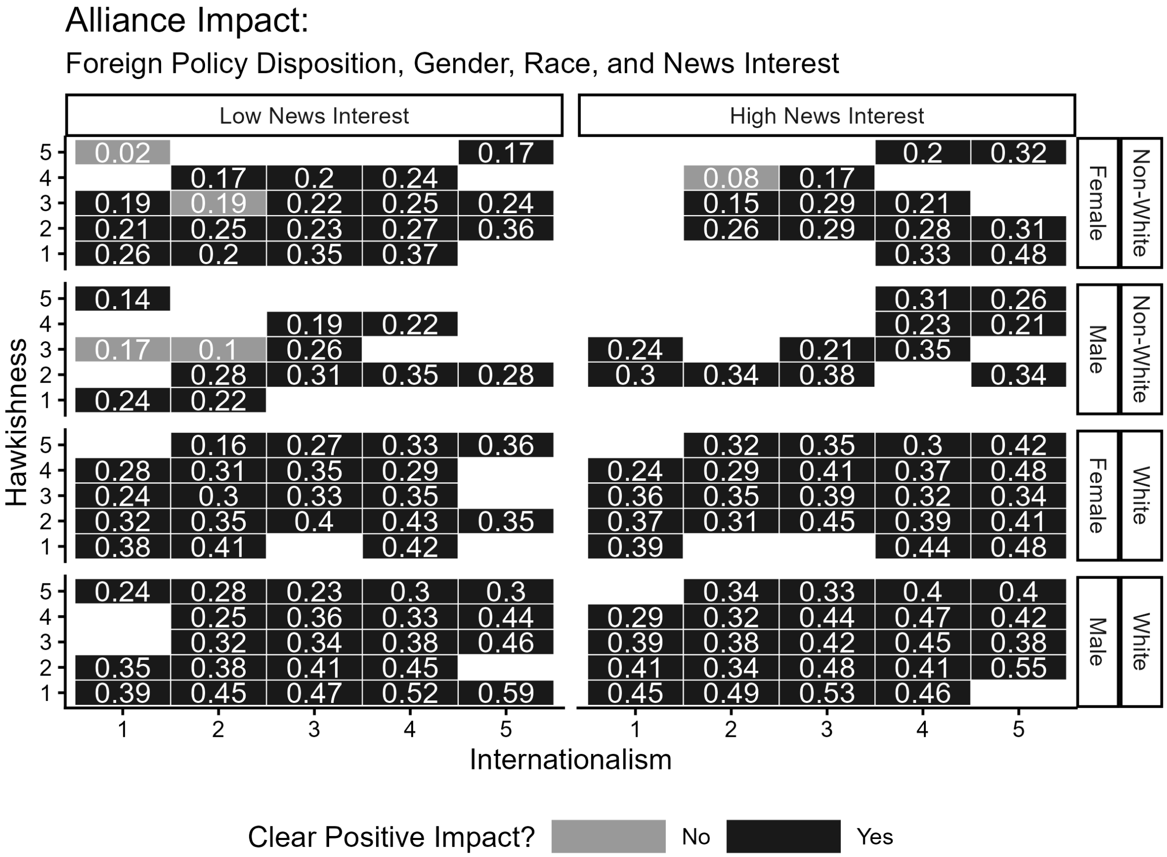

Figure 8. *Estimated impact of alliance on support for military intervention in different subgroups of respondents. The shaded area in each tile shows whether the median estimated treatment effect was clearly positive, and the text gives the exact median estimate.*

from low to high hawkishness.

The weakest impact of alliances, including some that cannot be distinguished from zero, occur at minimal internationalism and low news interest. Non-white females with these views are especially unlikely to respond to alliances. Otherwise, most demographic groups respond somewhat to TW's alliance treatment, but how much depends on their interest in the news and blanket support for international engagement. Alliances are unlikely to decrease support for intervention, at least in this experiment, and that is itself an important consideration.

These results show some potential uses of the hierarchical approach to heterogeneous effects.[11] Regularization moderates what might otherwise be extreme inferences about experimental subgroups. It also provides useful inferences about treatment heterogeneity that account for overlap among different sources of heterogeneity.

6 Conclusion

This note paper 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 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 perhaps two modifiers, relying on simple interaction tools is best. Machine learning is best for discovery of complex heterogeneity. When there are two or more theoretically informed modifiers, hierarchical modeling allows flexible and interpretable estimation of effect variation.

As a result, hierarchical modeling complements existing tools and should not replace them.

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

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, Carlisle Rainey and Rod Sturdivant 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. 2025. "Quantitative Political Science Research Is Greatly Underpowered." *The Journal of Politics* . 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.

DiGiuseppe, Matthew and Paul Poast. 2018. "Arms versus Democratic Allies." *British Journal of Political Science* 48(4):981–1003.

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.

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.

Martin, Lisa L. 2005. "The president and international commitments: Treaties as signaling devices." *Presidential Studies Quarterly* 35(3):440–465.

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.