

# Using Hierarchical Models to Estimate Heterogeneous Effects

Joshua Alley  
Assistant Professor  
Baylor University  
Joshua\_Alley@baylor.edu

February 10, 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.

Whether in observational or experimental studies, every independent variable social scientists examine impacts some units differently than others. Common estimands aggregate heterogeneous effects.<sup>1</sup> 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, international relations 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.

---

<sup>1</sup>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. 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 many potential modifiers, modeling such theories with interactions is not easy. The interpretation benefits of interactions diminish as researchers add more variables. Adding variables to an interaction further raises the risk of spurious inferences due to power concerns and picking up noise in ever smaller 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.<sup>2</sup> These tools can be difficult to interpret and implement, however, especially in smaller social science datasets. The relative lack of theoretical guidance and interpretability is especially problematic for testing multiple conditional arguments.

---

<sup>2</sup>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 power shortcomings of interactions 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. 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.<sup>3</sup>

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.

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

---

<sup>3</sup>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.

barrier to employing Bayesian methods, fitting a wide range of hierarchical models is straightforward with the `brms` package in R (Bürkner, 2017).<sup>4</sup>

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.

## 1 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 a regression specification and requires the same sort of care. Researchers need to define what variation is most important, link heterogeneous effects to theory, and structure modeling. This is especially important for pre-registration, and could reduce the amount of exploratory analyses in registrations. Not thinking carefully about 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

---

<sup>4</sup>I provide example code below and in the appendix.

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 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 papers do not have adequate power for main effects (Arel-Bundock et al., 2025), however, let alone small subgroups that many times have 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.<sup>5</sup>

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 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

---

<sup>5</sup>Extrapolations to a representative sample of a subpopulation might require poststratification.

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, and includes the treatment effects.<sup>6</sup>

The other regression models the treatment effects. In this regression, theory should inform which modifiers predict the treatment effect. To regularize the estimates, this second equation must include varying intercepts for groups within the data. Groups can be based on combinations of discrete modifiers and binned continuous modifiers. To maximize variation and preserve regularization, researchers should use continuous modifiers directly in the regression and bin them for setting groups. Not binning modifiers will create many small groups, leading the hierarchical component to lose value. As a result, this model includes both systematic and random variation in the impact of a key independent variable on different groups.

Setting groups is a critical task that requires balancing detail against the value of regularization. Larger groups can deviate more from the overall mean, as their data contributes more the posterior than the prior. A few small groups means less need for a hierarchical model. Smaller groups will be shrunk more towards the overall mean. Too many small groups and the regularization component will have less information to work with.

I now briefly describe a generic hierarchical model. For ease of exposition, consider making between-unit comparisons based on an experimental treatment. Start with  $N$  units indexed by

---

<sup>6</sup>If other units such as states define the groups, rather than combinations of modifying variables, then adding group-level predictors to this equation is essential. For example, 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.

$i$ , some of which receive a binary treatment  $T$ . Assume that the outcome variable  $y$  is normally distributed with mean  $\mu_i$  and standard deviation  $\sigma$ .<sup>7</sup>  $g$  indexes researcher-defined groups based on combinations of the modifying variables.

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

$$y_i \sim N(\mu_i, \sigma) \quad (\text{Likelihood})$$

$$\mu_i = \alpha + \lambda_g T + \mathbf{X}\beta \quad (\text{Outcome Equation})$$

$$\lambda_g = \theta_g + \mathbf{Z}\gamma \quad (\text{Heterogeneity Equation})$$

$$\theta_g \sim N(\mu_\theta, \sigma_\theta) \quad (\text{Individual Varying Intercepts})$$

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  $\gamma$ .  $Z$  can mirror any regression specification researchers might use for an outcome; linear combinations of variables, interactions, smooth functions. Interactions of the modifiers could capture processes where combinations of modifiers produce non-additive jumps in heterogeneity, for instance.

The random component of the heterogeneity equation is a series of individual-specific varying intercepts  $\theta_i$ . These are critical, because they capture group deviations from the systematic trends expressed in the design matrix  $\mathbf{Z}$ . Group outliers in the  $\lambda$  estimates are partially pooled back towards the overall mean  $\mu_\theta$ . The variance parameter  $\sigma_\theta$  controls the dispersion

---

<sup>7</sup>Researchers can and should use binary, categorical and other outcome likelihoods.

of the group effects around the overall mean.<sup>8</sup>

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  $\alpha$ ,  $\beta$ , and  $\gamma$  parameters is especially helpful. Priors also help regularize estimates by pulling extreme groups towards the overall mean. Working with posterior distributions also provides a wealth of information about effect heterogeneity and propagates uncertainty. Last, Bayesian models can provide useful diagnostics such as divergent transitions that can be warnings of specification problems. A model that is hard to fit oftentimes should be modified, not brute-forced through the sampler.

In interpreting these estimates, researchers should leverage the full range of information from the different parameters. First, the  $\lambda$  posteriors give the impact of the treatment on each individual, and are the core quantity of interest. All  $\lambda$ s reflect a systematic component from the predictors in  $Z\gamma$  and a random variations from  $\theta$ .  $\gamma$  parameters can, depending on the regression, be interpreted as the impact of a change in a modifier on the treatment effects. For example, a  $\gamma$  of .1 on a binary modifier means that  $\lambda$  is .1 higher in expectation when the modifier is one, and .1 lower when it is zero.  $\sigma_\theta$  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.

## 2 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

---

<sup>8</sup>In brms, using non-linear syntax can express a model with a treatment, two controls, and three modifiers as:  $y \sim \text{lambda} * \text{treat} + \text{controls}$ ,  $\text{lambda} \sim \text{mod1} + \text{mod2} + \text{mod3} + (1 | \text{mod1} + \text{mod2} + \text{mod3})$ ,  $\text{controls} \sim \text{control1} + \text{control2}$ ,  $\text{nl} = \text{TRUE}$

interpretable framework that encompasses many potential modifiers. This helps examine theories that predict how an effect varies and compare sources of variation.<sup>9</sup> 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 are themselves informative. Hard to fit models may need to be simplified, especially if there are many small groups. Again, many small groups will also reduce the regularization benefits of the hierarchical model, as there is less information to estimate  $\sigma_\theta$ .

Furthermore, hierarchical models can show general trends, but will not make powerful comparisons between every group treatment effect. Researchers who want to compare specific effects 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. Interpreting triple interactions between a variable and two modifiers is challenging. The advantages of hierarchical modeling increase with the number of modifiers until too many modifiers create many

---

<sup>9</sup>Rescaling variables in the heterogeneous effects equation can aid model fitting and coefficient comparisons (Gelman, 2008).

small 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, machine-learning is a better tool.

Data size is relevant to model selection as well. All of these models benefit from more data, as machine-learning models can draw out patterns and interactions have greater power. Hierarchical models also benefit from more data, but are less sensitive to outliers in small samples. The challenge is that some hierarchical models can take a long time to fit on large datasets. Variational approximation can ameliorate this problem, but these methods require careful validation.

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

Characteristic	Hierarchical Models	Interactions	Machine Learning
Factors	Two or more	One or two	Many
Ideal Sample Size	Medium	Medium to large	Large
Complexity	Medium	Low	High
Computational Cost	Model dependent	Low	Variable
Interpretability	High	High	Low
Modifiers	Specified	Specified	Discovered/specifyed

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. Hierarchical modeling is especially important when there are multiple modifiers from different

theories of treatment heterogeneity or heterogeneous effects.

Table 1 summarizes some relevant characteristics of hierarchical, interaction and machine learning approaches to heterogeneous effects. Interactions work well with a few modifiers, a medium to large sample, and are easy to integrate with theory. Machine-learning tools are best for discovery with many modifiers in large datasets, but can be harder to tie into existing arguments. Hierarchical models are marginally more difficult to use than interactions, but researchers gain an interpretable way to tackle multiple modifiers in a theoretically-driven model. I now document gains of hierarchical models compared to interactions with simulated data. I do not include machine-learning tools in the simulation because they often produce different effects, such as individual-level estimates that are not directly comparable to the group-level hierarchical and interaction estimates.

### 3 Performance on Simulated Data

To assess how this hierarchical model compares to interactions, I first assess their performance on simulated data. Each simulation approximates the two most common applications of these models. The first simulates estimating treatment heterogeneity 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 of heterogeneous effects. In both simulations, I attempt to push the limits of group construction to illustrate when hierarchical modeling becomes more or less accurate.

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 outcome 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). Lower 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. Fully crossed interactions are also a standard tool for estimating effects in each combination of experimental treatments. In the hierarchical model, I use fully crossed interactions of the group variables in the heterogeneity equation and set a random intercept for each group.

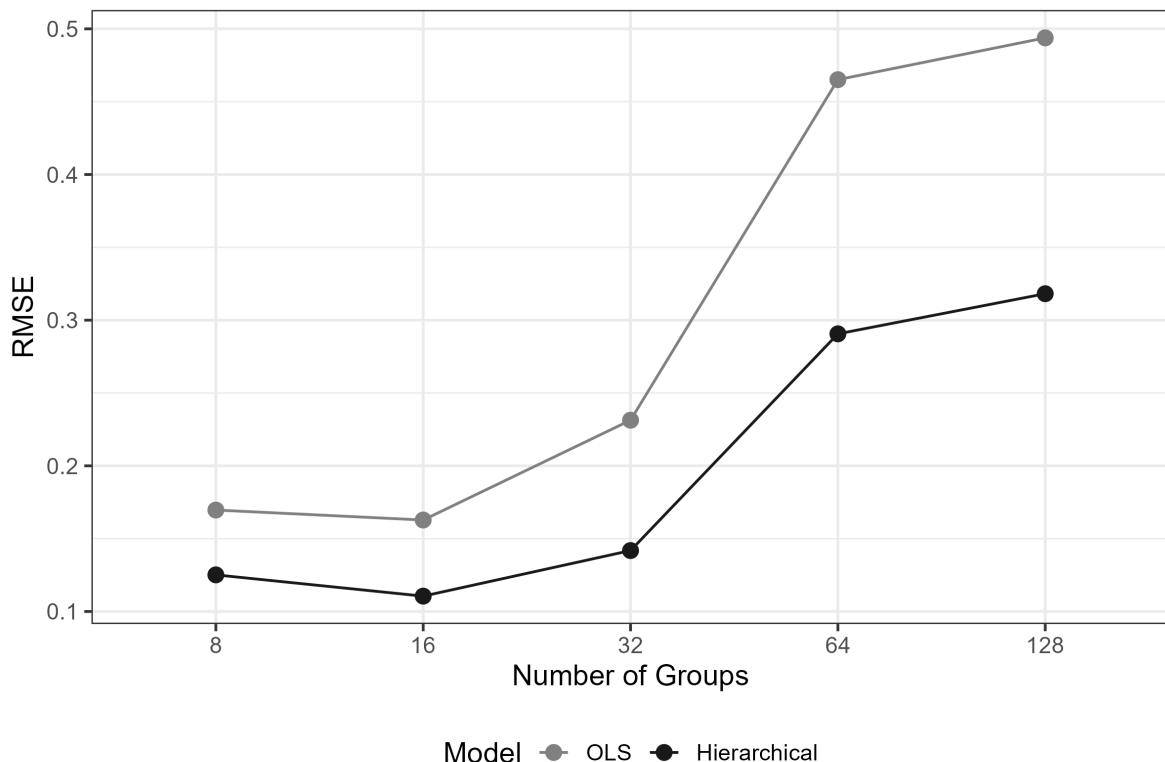
Figure 1 presents the results of this first simulation. When the groups are large because there are three grouping variables, there is a small gap in model performance. The hierarchical model is marginally more accurate, however. But as group numbers rise and the number of observations in each group 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.

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 clearly off. This means that even with slight bias for all groups, the hierarchical

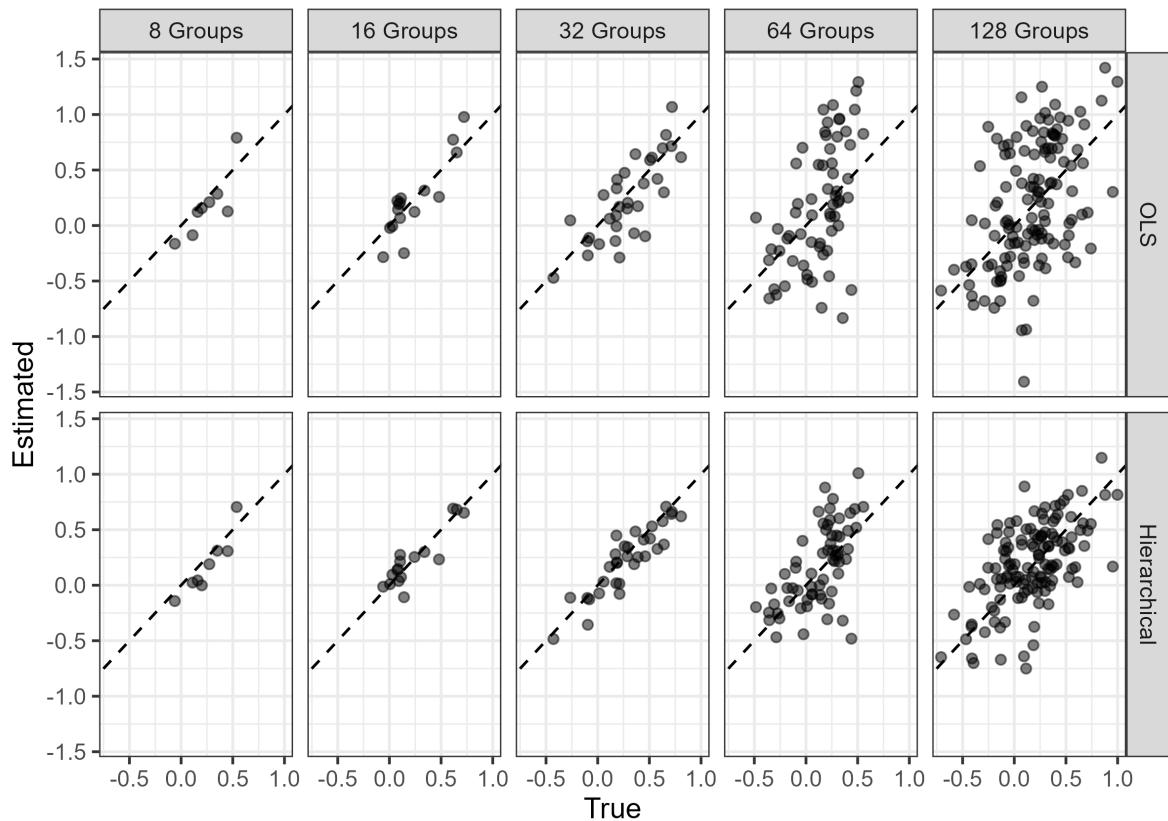
### Equal Group Size: RMSE of Group Treatment Effect Estimates

$n = 2000$



**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.

### Equal Group Size: Estimated vs. True Effects



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

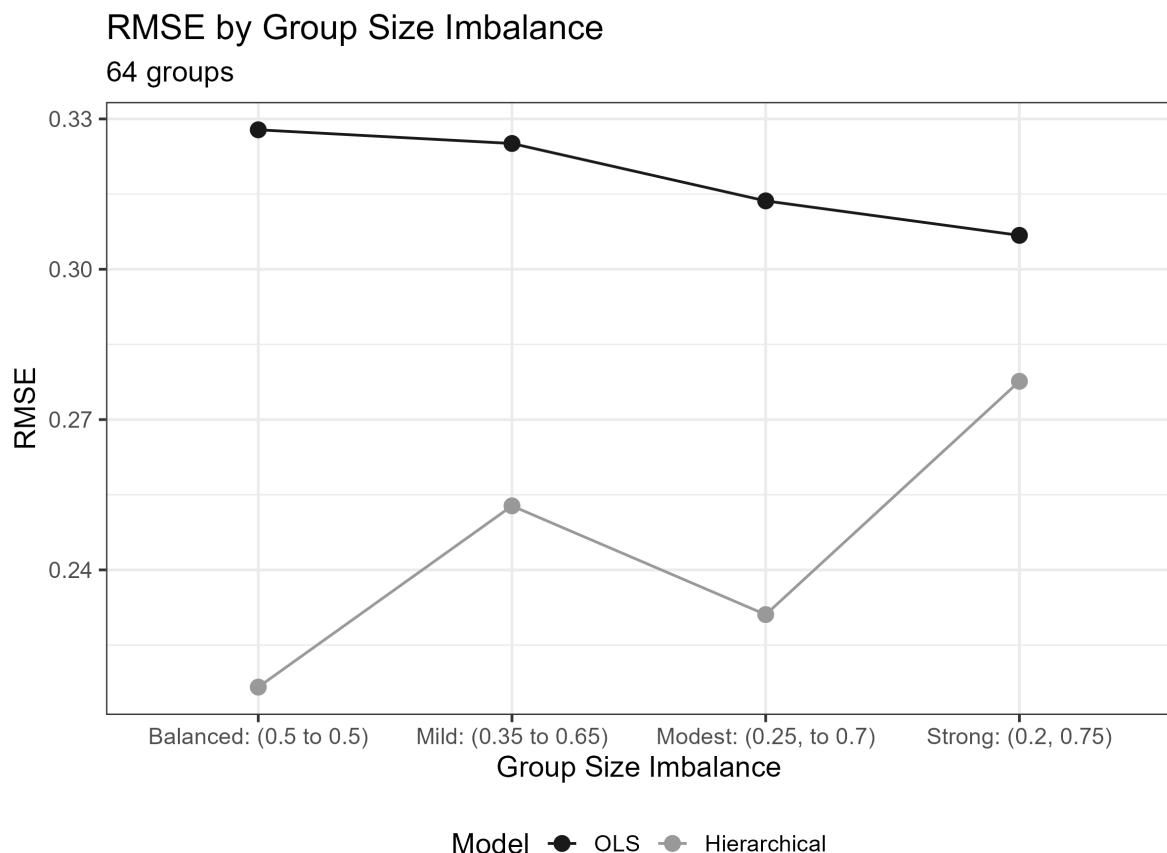
estimates have lower overall root mean squared error. As the number of groups increases and group size falls, the estimates for both models become more noisy, but the hierarchical model estimates are far less variable. Lower variance makes the hierarchical model much more accurate on average.

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 with unbalanced group sizes. The relative benefits of hierarchical modeling fall as the number of small groups in the data rises, however. 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.

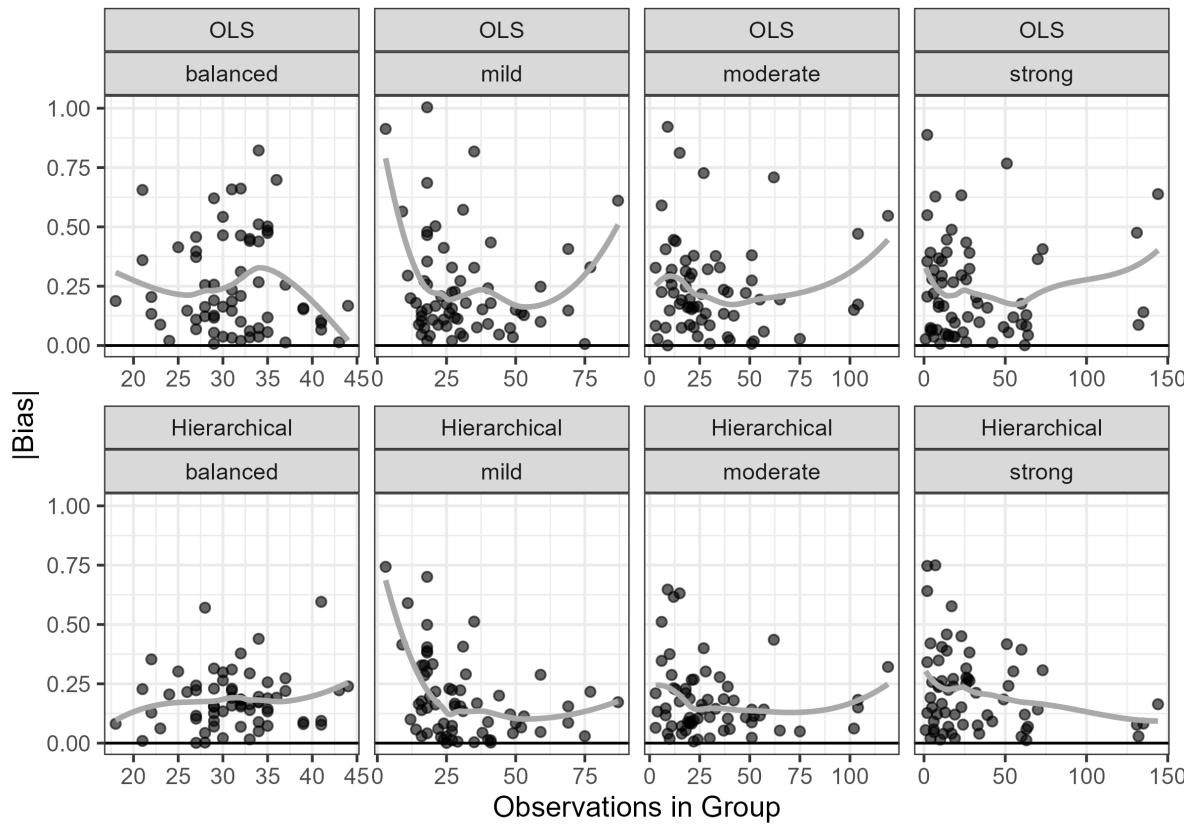
Under balanced groups, the hierarchical model has a significant advantage. The smaller and more numerous the groups, the less valuable hierarchical regularization becomes, however. Small groups add far less information and are strongly pulled to the overall mean, which itself is a function of the larger 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 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



**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.

### Absolute Bias by Group Size



**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.

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 more data and less bias. This is partial pooling in action. Small groups do not have much information, so the prior informs those estimates more than the limited data. Large groups have enough data to deviate from the prior and hit the true value.

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 the estimate will often be more accurate than interactions regardless. Individual estimates may not be perfectly unbiased and comparing groups will be difficult.

## 4 Example Application: Alliances and Public Support for War

In the following, I demonstrate how the hierarchical approach works in practice and compares to prior applied work 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 when the beneficiary of that intervention is a U.S. ally. In a factorial experiment with vignettes, they find a 33% average increase in support for military intervention on behalf of allies, compared to non-allies. 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 many 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 change some conclusions by pooling noisy estimates in small groups. 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.

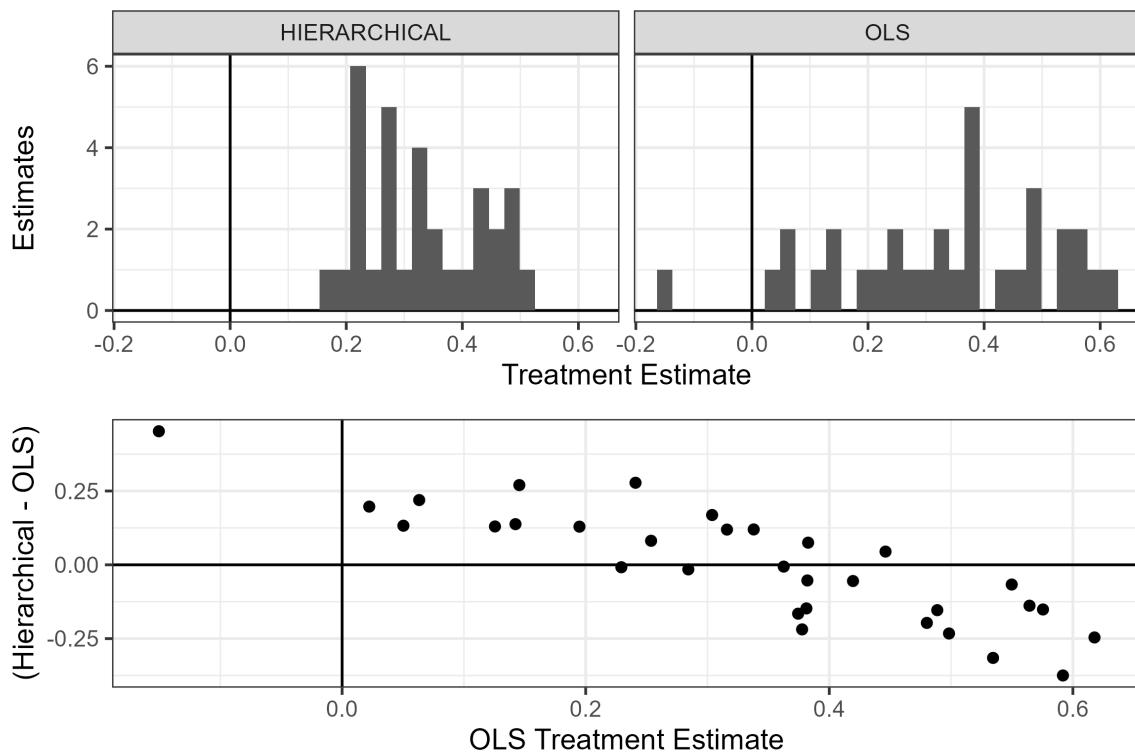
#### *4.1 Differences by Experimental Scenario*

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 and group-specific intercepts. This mirrors TW's model, but regularizes the estimates with partial pooling.

Figure 5 compares the estimated alliance treatment effects across the experimental groups with the OLS and hierarchical models. This figure illustrates the regularization of hierarchical modeling in two ways. It shows the difference between the median hierarchical estimate and the OLS estimate for each combination of treatment groups and plots the distribution of effects.

First, the hierarchical estimates are much less variable and more tightly clustered around the overall mean. This occurs because the OLS model mechanically forces individual treatment effect estimates to the value of the coefficients and corresponding subgroup means. Given the

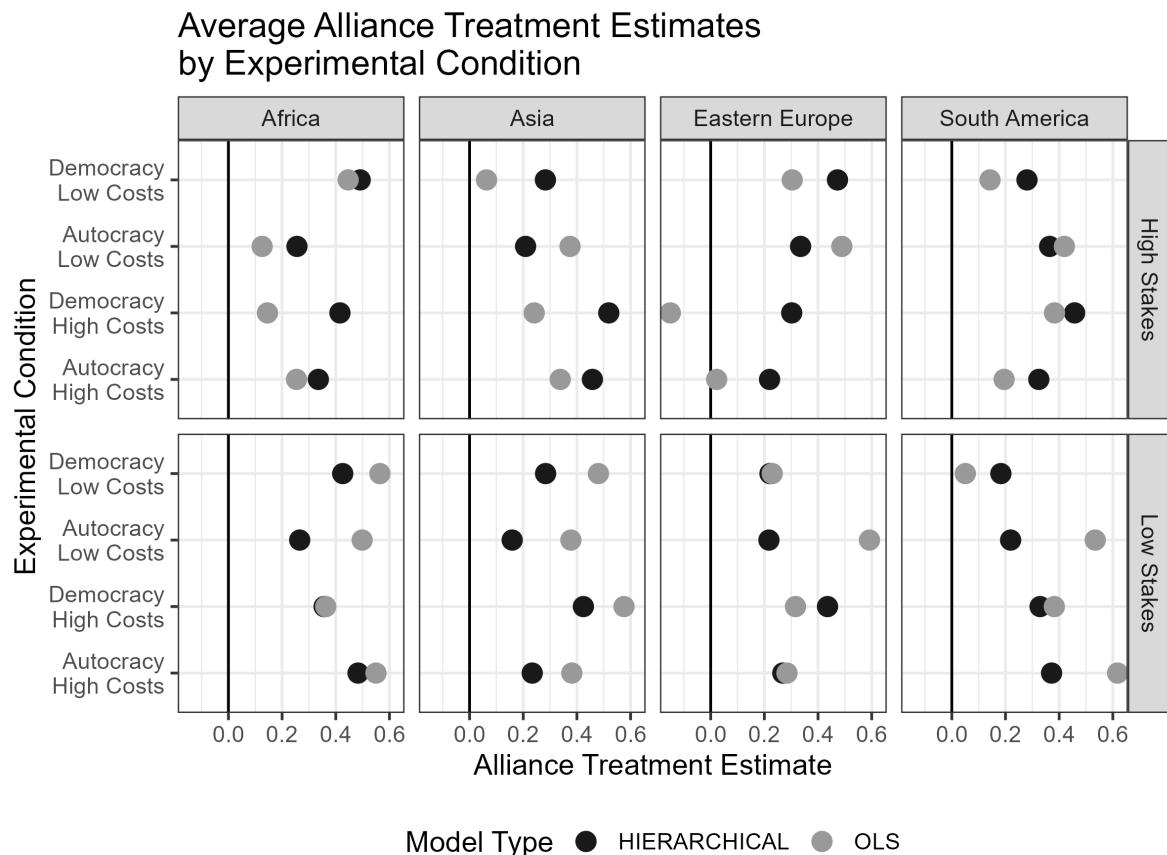
## Comparison of Heterogeneous Treatment Estimates Heterogeneity from Experimental Conditions



**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.

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 hierarchical and the OLS estimate for 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 values. The OLS model finds one scenario where alliances reduce support by almost 20%, but the hierarchical model treats this as noise and brings it towards the grand mean.



**Figure 6.** Comparison of OLS and hierarchical estimates of the impact of alliance across experimental conditions. Each point gives the effect estimate for each group.

I compare the estimates for each group with greater detail about what shapes each heterogeneous treatment in Figure 6. This clearly shows the reduced variation in the hierarchical estimates across the different experimental conditions. In some cases, the hierarchical and OLS estimates are very similar, but the model suggests that in a smaller sample, saturated interactions of treatment effects overstate effect heterogeneity. For some conditions, such as a low-stakes, low-cost intervention to support a South American autocracy, the hierarchical estimate is 30% lower than the OLS estimate. The condition where TW find a 20% drop in support for an ally compared to a non-ally is a high-cost intervention on behalf of an Eastern European democracy. That would be concerning, but the hierarchical model suggests that this is closer to 30%, which matches findings on how randomly assigning NATO membership to different Eastern European states impacts U.S. public support for war (Tomz, Weeks and Bansak, 2023).

Given the very large average effect of alliances, these results 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. The hierarchical models suggest that low stakes are not as influential as the OLS results indicate, for example. For other, smaller effects, regularization with a hierarchical model might change inferences in subgroup analysis.

## ***4.2 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, and gender to modify the impact of alliances. I selected these variables because Tomz and Weeks examine party and political interest as potential modifiers, but gender (Barnhart et al., 2020) and race are also salient modifiers. I control for other experimental manipulations, age and education. Age and education are important group-level controls because they are group-level variables that are correlated across party, race and gender. Following TW's OLS analysis,

I use a Gaussian likelihood, although the outcome is a binary variable.

The resulting hierarchical model thus encompasses four subgroup comparisons in TW's appendix while adding two more modifiers. All of the comparisons encapsulated in the hierarchical model would require at least six pairwise interaction models. What TW present in four figures, I am able to do in two while adding more information.

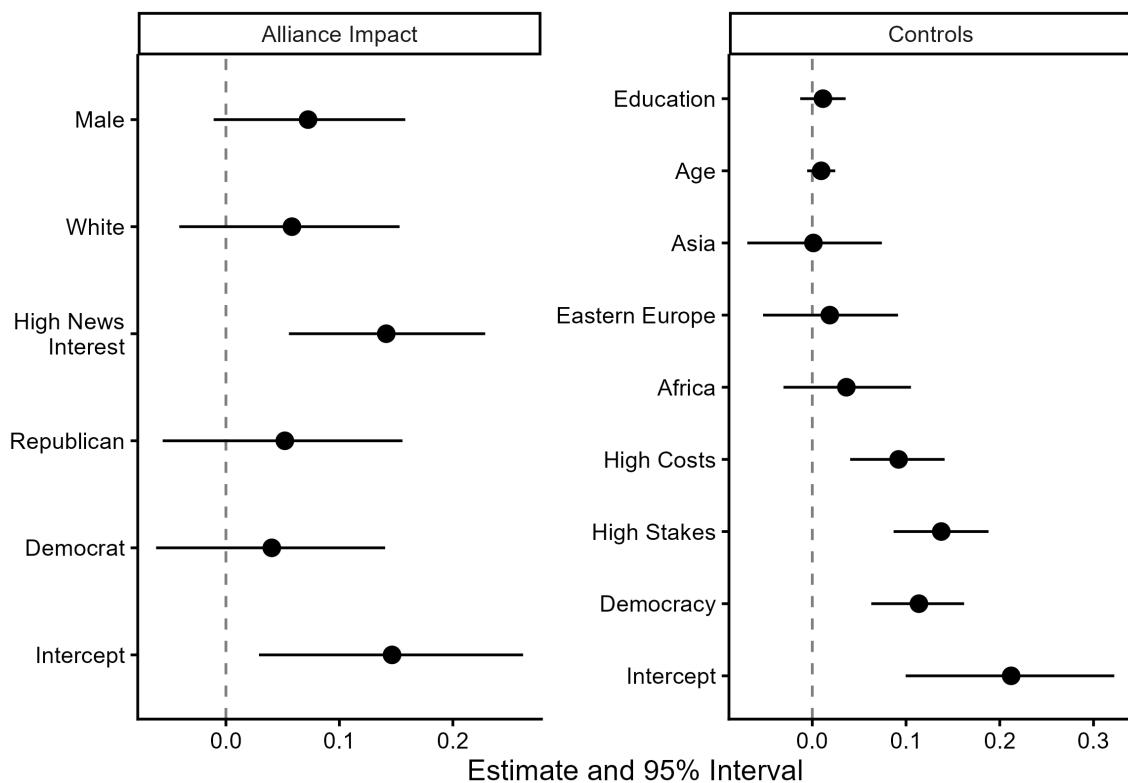
To start, I plot the two regressions of hierarchical model. Figure 7 plots how different variables shape either the impact of an alliance or support for intervention. One facet of this figure gives the coefficient estimates from the model of heterogeneity, and a second gives the control estimates.

High news interest is the strongest predictor of how alliances impact support for war. Respondents with high news interest respond to alliances by .15 more than others in this survey, and that relationship could be as small as .05 or as large as .22. 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.

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 is theoretically meaningful, as it captures the impact of alliances when all the modifiers are zero. For independent, non-white women with low news interest, alliances increase support for force by roughly .15.

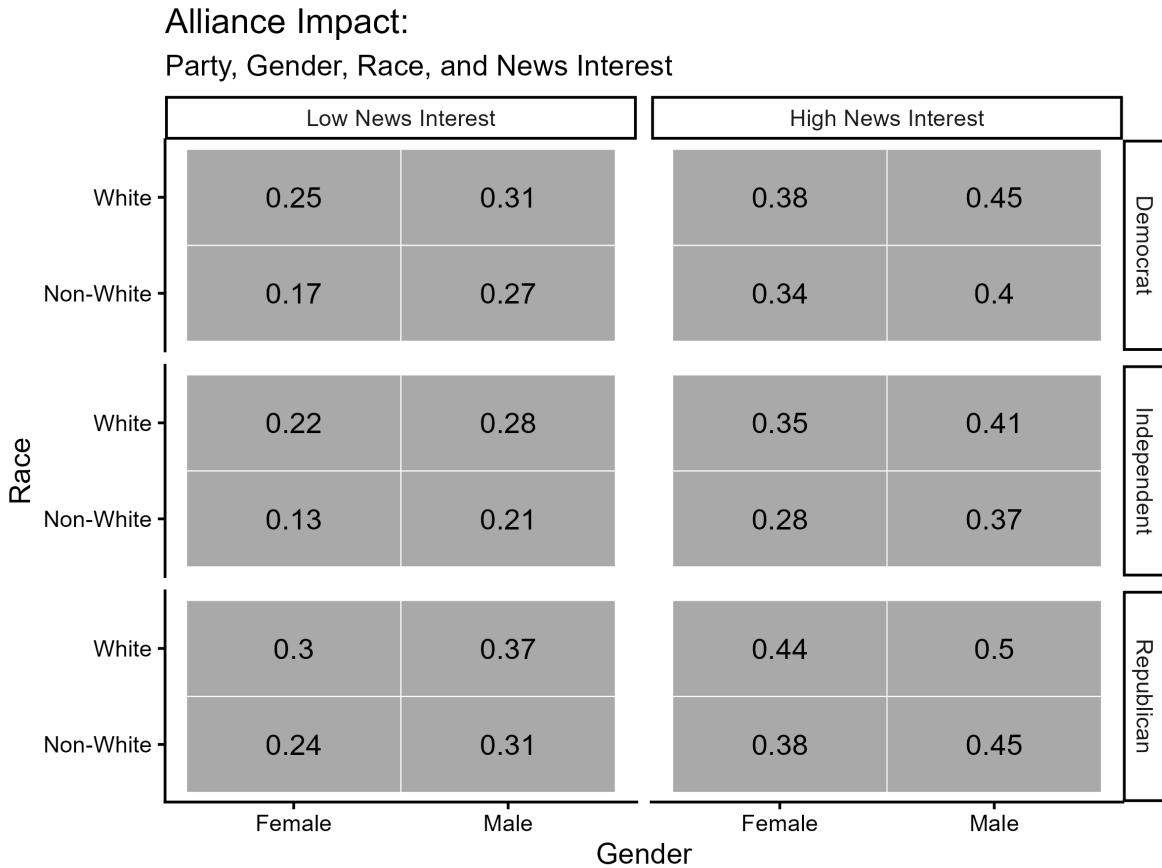
The control equation suggests that high stakes, high costs, and partner democracy all increase support for intervention. Region is less consequential, as neither Asia, Eastern Europe

## Determinants of Support for Force and Alliance Impact Hierarchical Model



**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.

nor Africa is clearly different from the reference category of Latin America. Because neither education nor age are equal to zero, the intercept is not directly interpretable.



**Figure 8.** Estimated impact of alliance on support for military intervention in different subgroups of respondents. Text gives the exact posterior median estimate.

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 partisanship. There are 24 unique estimates, all of which are clearly greater than zero. Alliances thus increase support for using force among all demographic groups, but the magnitude of the effect varies.

Alliances most impact white Republican men with high interest in the news. In this group,

the impact of alliances is .5, because this group was skeptical of intervention without an alliance rationale. The gap between high and low news interest in the other demographic groups is .15, which matches the coefficient in the heterogeneity equation.

The weakest impact of alliances appears for non-white, politically independent women with low news interest. Here, alliances increase support for force by 13%. White respondents respond by 8% more, Democrats by 4% more, and Republicans by 10%. The accumulation of these different effects create substantial gaps between the groups.

These results show some potential uses of the hierarchical approach to heterogeneous effects.<sup>10</sup> 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.

## 5 Conclusion

This paper details why, when and how 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.

---

<sup>10</sup>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 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.
- Tomz, Michael, Jessica LP Weeks and Kirk Bansak. 2023. “How membership in the North Atlantic Treaty Organization transforms public support for war.” *PNAS Nexus* 2(7):pgad206.
- 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.