

Appendix: Alliance Participation, Treaty Depth, and Military Spending

This online appendix provides more detail about the multilevel model and checks the results. I also briefly describe a single-level test of the depth hypothesis and assess other measures of alliance treaty depth.

1 Priors

Table 1 summarizes the prior distributions in the multilevel model. All priors are weakly informative relative to the scale of the data. ν is the degrees of freedom for the t-distribution, and the gamma prior is the recommended prior for STAN (Juárez and Steel, 2010).

2 Hamiltonian Monte Carlo Diagnostics

There were no divergent iterations in either sample running 4 chains for 2,000 iterations with 1,000 warmup iterations. The \hat{R} is less than 1.1 for all parameters in both samples. Trace plots in Figure 1 indicate good mixing of the chains for the alliance-level parameters. Taken together, all of this implies that the chains adequately explored the posterior distribution.

$$\begin{aligned} p(\alpha) &\sim N(0, 1) \\ p(\sigma) &\sim \text{half-}N(0, 1) \\ p(\alpha^{yr}) &\sim N(0, \sigma^{yr}) \\ p(\sigma^{yr}) &\sim N(0, 1) \\ p(\alpha^{st}) &\sim N(0, \sigma^{st}) \\ p(\sigma^{st}) &\sim \text{half-}N(0, .5) \\ p(\sigma^{all}) &\sim \text{half-}N(0, .5) \\ p(\beta) &\sim N(0, .5) \\ p(\gamma) &\sim N(0, .5) \\ p(\nu) &\sim \text{Gamma}(2, 0.1) \end{aligned}$$

Table 1: Summary of Priors in Multilevel Model

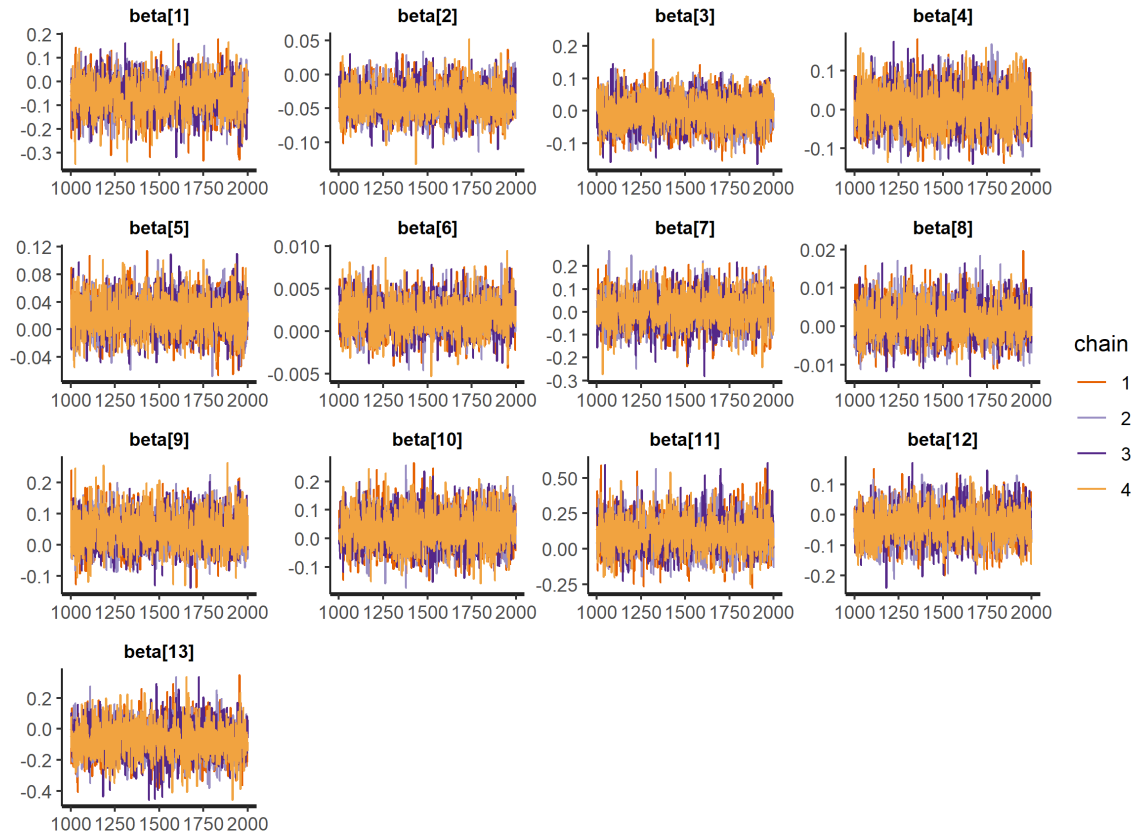


Figure 1: Traceplot of alliance level parameters in the non-major power sample.

3 Fake Data Simulation Check

With any complicated model, simulating fake data and seeing if the model can recover known parameters is essential. Fake-data simulation helps validate results from observed data and identify problems. This section summarizes results from fitting the multilevel model to fake data.

I simulated a dataset of 2000 t-distributed observations with 50 states observed for 200 years and 100 alliances. The outcome has a different scale than the military spending outcome variable, so coefficient values here do not match the paper. I then simulated two state and alliance level variables and took a piece of the matrix of state membership in alliances. Last, I ran the model without evaluating the likelihood, generating a posterior prediction of the outcome based on the fake data.

To check whether the model could recover known parameters, I took the 12th draw of the posterior distribution. This draw included a simulated outcome for each observation and a set of coefficients. I then fit the multilevel model on the simulated outcome values and checked whether the credible intervals contained the corresponding parameter values. If a parameter is within the 90% credible interval, the model captures it.

The model recovers known parameters with a high degree of accuracy. The two credible intervals of the alliance-level regression include the known values. Credible interval coverage for the variance hyperparameters and γ parameters is also acceptable. Among the λ parameters, 93 of 100 intervals contain the known λ value. Given the large number of parameters and smaller sample, this is acceptable accuracy. Even the seven inaccurate confidence intervals were quite close— all were within .015 of the known parameter.¹

In summary, convergence diagnostics and fake data fitting both suggest that the multilevel model gives valid estimates. No convergence diagnostics indicate problems exploring the posterior. Just as importantly, the model can recover known parameters from fake data.

4 Robustness Check: Single-Level Regression

Though the multilevel model best reflects the theory, I also fit some more standard panel data models. In what follows, I briefly present results from robust regressions of state-year percentage changes in military spending in the same sample of non-major powers. As in the multilevel model, I applied the inverse hyperbolic sine transformation to the outcome. In these models, I employ two indicators of alliance depth. The first is the average depth of a state's alliances. The second is a dummy which equals 1 if a state has at least one alliance with greater than average depth. Both variables compare states with different depth in their alliance portfolio. In addition to the state-level controls in the multilevel model, I included average alliance size, average allied democracy and the log of total allied capability as controls.

I estimated several models, including robust regressions on non-major powers and non-major powers in alliances. Comparing non-major powers with at least one alliance provides a crude approximation of the depth coefficient in the multilevel model, which compares deep and shallow

¹Fine margins around these intervals implies that the exact number of accurate λ intervals is sensitive to simulation variance.

alliances. I also applied state and year fixed effects to an OLS model of percentage changes in defense expenditures. The estimated association between average treaty depth and military spending changes is summarized in Figure 2. Results are inconsistent- I do not reject the null hypothesis for the average depth measure unless the model includes state and year fixed effects. The deep alliance dummy coefficient estimate is negative and statistically significant across all samples and model specifications, however. The dummy variable is closest to DiGiuseppe and Poast (2016)'s design.

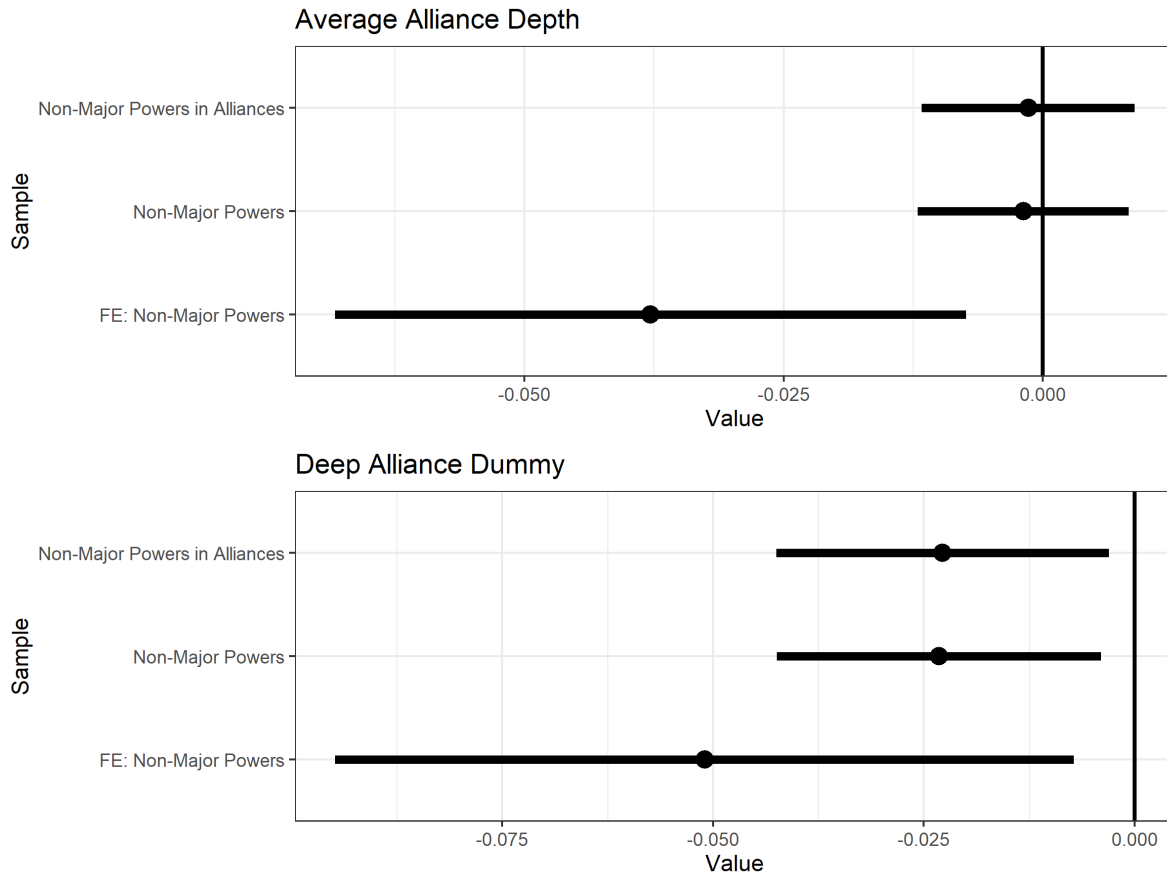


Figure 2: Estimated effect of average alliance treaty depth or a dummy indicator of participation in a deep alliance on percentage changes in non-major power military spending from 1816 to 2007.

5 Alternative Measures of Alliance Treaty Depth

This part of the appendix compares my latent measure of treaty depth with similar measures in earlier research. I first show that there are important differences between my measure and the military institutionalization measure of Leeds and Anac (2005), but using military institutionalization instead of latent depth generates similar inferences about how treaty depth modifies the impact of alliance participation. Then I describe the conceptual and empirical differences between my latent measure and that of Benson and Clinton (2016).

5.1 Leeds and Anac 2005

Leeds and Anac (2005) create an ordinal measure of alliance treaty depth to study whether alliance institutionalization improves treaty reliability and performance in offensive and defense alliances. They argue that commitments of an integrated military command, common defense policy, or any basing rights generate high military institutionalization. Official contact between military officials, formal organizations, providing training or technology, subordination of forces, or specific contributions reflect moderate institutionalization. If at least one factor is present, Leeds and Anac assign alliance the highest corresponding level of institutionalization.

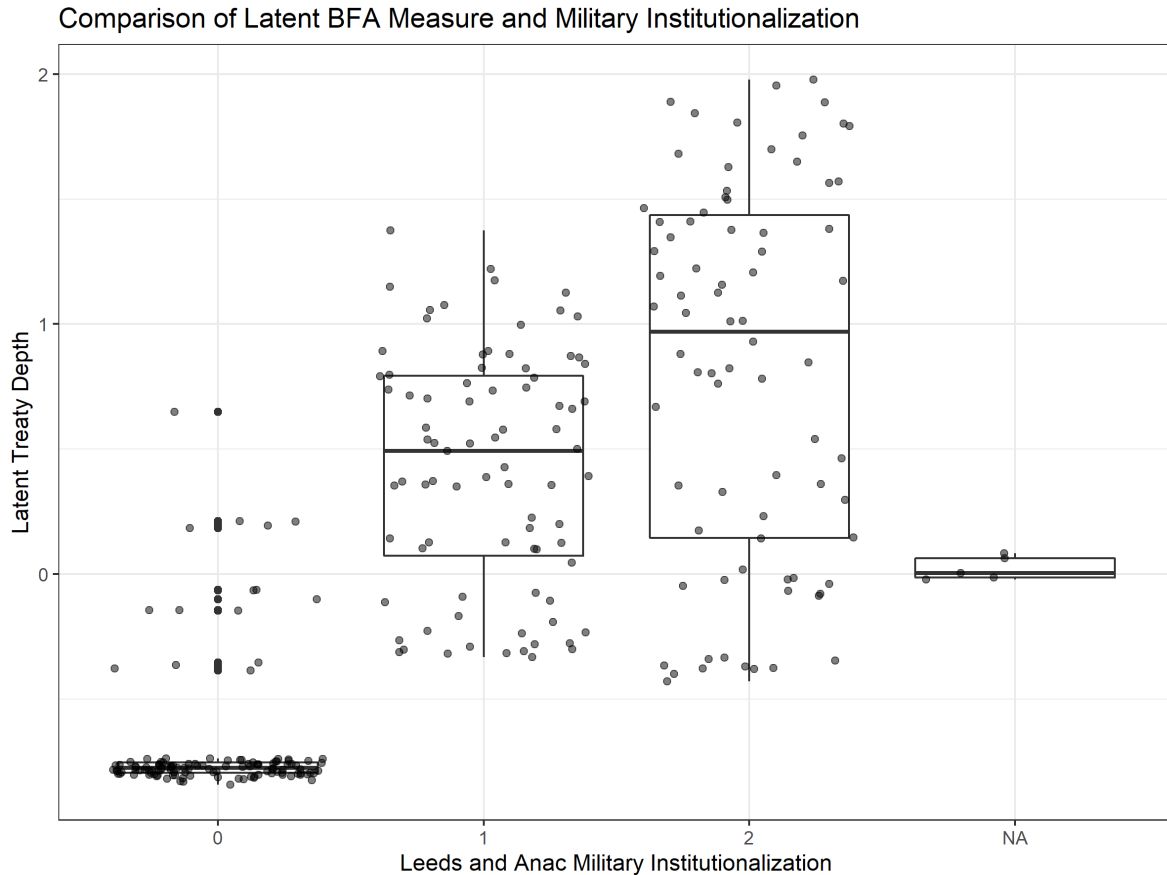


Figure 3: Scatter plot of latent treaty depth across the values of military institutionalization from Leeds and Anac (2005). The box plots summarize the distribution of latent treaty depth within each category of military institutionalization. Points are jittered within each level of the institutionalization score.

This approach assumes that alliances with multiple sources of depth as just as deep/institutionalized as alliances with one factor and understates the amount of variation in alliance treaty depth. Even so, as Figure 3 shows, this ordinal measure and my latent measure are positively correlated. The deepest alliances on the latent measure also have the highest military institutionalization score, because they rely on similar variables. There are substantial differences within each category and

overlap in the latent scores across the categories, however. For example, some alliances that Leeds and Anac (2005) assign a moderate institutionalization score have more depth than alliances with high institutionalization scores because these alliance treaties contain multiple sources of depth.

There are two sets of alliances where my measure makes a marked departure from Leeds and Anac. First, there are some alliances with no institutionalization that my measure assigns some depth to. This difference is the result of companion military agreements, which I include as a source of depth in addition to Leeds and Anac's variables. Second, Leeds and Anac assign missing values to some institutionalization scores if all sources of high or moderate depth are missing, which is a reasonable choice with their measurement strategy. My latent measure gives these treaties some depth, because it accommodates missing data on a subset of variables.

Although I the latent depth measure has some advantages, I find similar results with the ordinal measure of military institutionalization. To check the robustness of my results, I implemented the same multilevel model of non-major power military spending, but replaced the mean latent treaty depth variable with the military institutionalization measure. Table 2 summarizes the results.

	mean	sd	5%	95%	n_eff	\hat{R}
Constant	0.003	0.058	-0.094	0.096	2688.743	1.001
Military Inst.	-0.035	0.024	-0.075	0.004	3960.366	1.000
Uncond. Milsup.	-0.018	0.042	-0.087	0.051	3450.201	1.000
Econ. Link	0.015	0.049	-0.066	0.096	3056.525	1.000
FP Conc.	0.025	0.025	-0.017	0.066	4115.104	1.000
Number Members	0.002	0.002	-0.001	0.004	3696.671	1.000
FP Similarity	-0.005	0.065	-0.110	0.105	2697.860	1.001
Democratic Membership	0.001	0.004	-0.006	0.008	3134.146	1.000
Wartime	0.047	0.051	-0.037	0.132	3879.252	1.000
Asymmetric	0.056	0.059	-0.037	0.155	2909.115	0.999
US. Mem	-0.033	0.049	-0.112	0.049	2617.047	1.000
USSR Mem.	-0.079	0.098	-0.237	0.083	3185.998	1.000
σ Alliances	0.143	0.054	0.060	0.234	914.843	1.003

Table 2: Results from an analysis that replaces the latent measure of treaty depth with an ordinal measure of military institutionalization from Leeds and Anac (2005). The negative correlation between military institutionalization and the impact of alliance participation on military spending matches earlier conclusions about the way treaty depth impacts military spending.

The same finding about treaty depth holds when the analysis uses military institutionalization in place of mean latent treaty depth. Military institutionalization and the impact of alliance participation on military spending are negatively correlated. 93% of the posterior probability in the depth coefficient is negative, which matches Hypothesis 3.

Figure 4 helps assess Hypotheses 1 and 2. Among alliances with no institutionalization, most treaties have a positive effect. Alliances with moderate institutionalization have mixed effects. Last, participation in alliances with high institutionalization tends to reduce military spending. This corresponds to the predictions of Hypotheses 1 and 2. The trend across military institution-

alization is less clear than with the latent measure of treaty depth, probably because the military institutionalization measure understates variation in treaty depth.

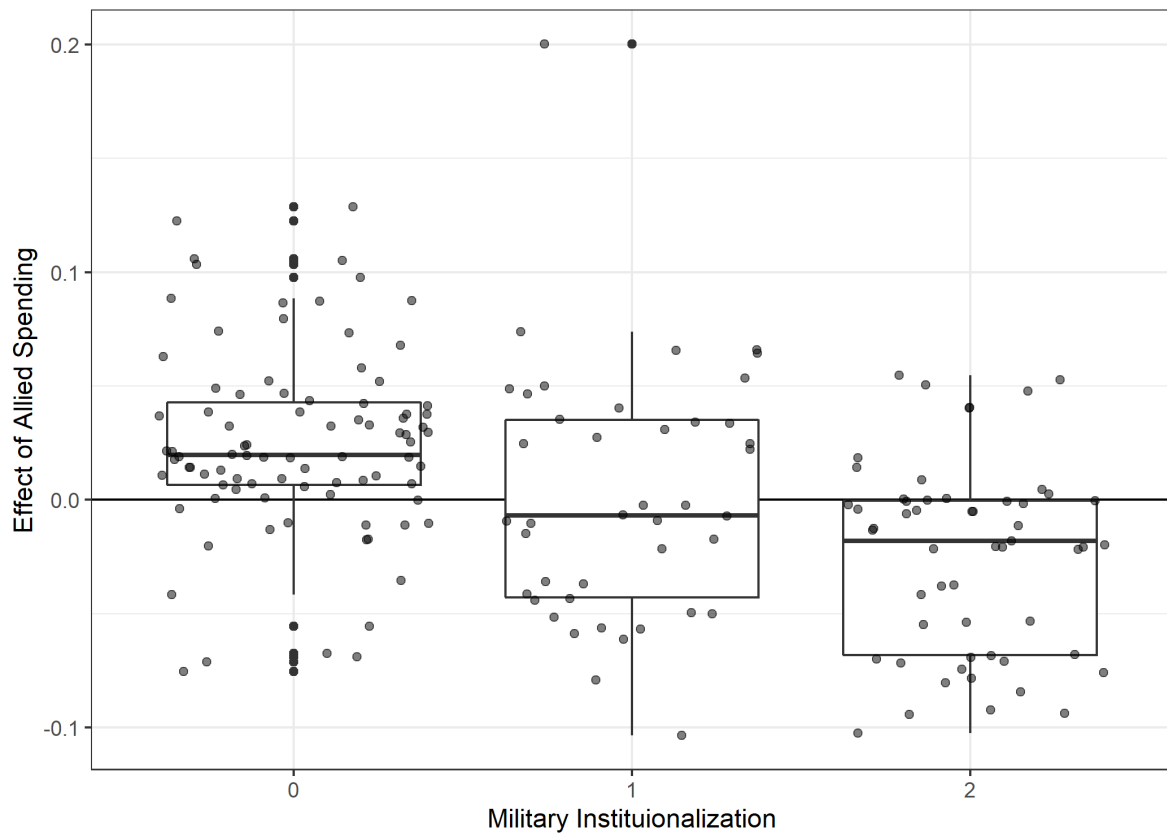


Figure 4: Scatter plot of λ parameters against the values of military institutionalization. The box plots summarize the distribution of points within in each military institutionalization value. All points are jittered to make the plot more legible.

5.2 Benson and Clinton 2016

Having established that an ordinal measure of treaty depth produces similar results, I now explain why I did not use an existing latent measure of treaty depth. Benson and Clinton (2016) use a latent variable model to estimate alliance scope, depth and capability. I do not use their measure of depth because it captures a different concept, and thus diverges from my aims in this project. Benson and Clinton define depth as the general costliness of alliance obligations, which leads them to include other variables and measure the depth of neutrality pacts. I define depth in terms of military coordination and cooperation and am only interested in alliances with active military support. My focus on defensive and offensive alliances follows existing scholarship on alliance participation and military spending. These conceptual differences, along with my use of

a different estimator, lead to different conclusions about the factor loadings and the latent depth scores.

Benson and Clinton aim for a broad measure of alliances, so they include neutrality pacts in their data. There is an understandable choice, but it means their measure diverges from the my argument, which focuses on alliances with military support. Neutrality pacts are qualitatively different, because peacetime coordination is not focused on ensuring the delivery of military support.

In Benson and Clinton's measure, neutrality pacts have very little depth, which is unsurprising. Neutrality is less costly in general. Only a few alliances with only neutrality obligations have any depth. In estimation, neutrality pacts are something like a reference category for military alliances. Neutrality pacts have limited depth, but they inform inferences about the depth of alliances with offensive or defensive promises. By including neutrality pacts, Benson and Clinton's model compares alliances with generally limited obligations to alliances with active military support.

Again, Benson and Clinton's decision to measure the depth of neutrality pacts is not a problem for their paper. It limits the applicability of their measure to this project, however. I do not address neutrality pacts in my argument, so my measure of depth emphasizes variation only among offensive and defensive alliances. My theoretical focus on military cooperation also leads me to exclude indicators of secrecy and economic aid that Benson and Clinton use to predict treaty depth. Although Benson and Clinton's measure is close to my purposes, it captures a slightly different concept.

Given conceptual differences and my decision to use an estimator that breaks problematic correlations between the latent variables and dependence structure (Murray et al., 2013), I draw different conclusions about the factor loadings and distribution of treaty depth. Figure 5 describes the key differences between my latent measure of depth and Benson and Clinton's measure. In the top panel, I look at differences in the factor loadings across the two models.² Benson and Clinton break the ATOP policy coordination variable into military contact and common defense policy, but I treat this as an ordinal variable, which is the largest source of depth.³ I also find larger correlations between formal organization and integrated military command and latent depth than Benson and Clinton. It is harder to distinguish between the other loadings, but Benson and Clinton's measure assigns marginally more weight to military aid and basing rights.

In the bottom panel of Figure 5, I plot my measure of treaty depth against Benson and Clinton's. To facilitate this comparison, I rescaled both depth measures by dividing them by one standard deviation. Benson and Clinton's measure suggests 18 alliances are 2 or more standard deviations from the mean, while my measure contains 9 such alliances. Many treaties with high depth on Benson and Clinton's measure have lower depth relative to other alliances on my measure. The two measures identify a common set of six extremely deep alliances, but disagree about how distinctive they are from other treaties. Salient distinctions in the relative depth of alliance treaties follow from differences in the factor loadings.

²Recall that I removed economic aid and secrecy from the observed data in my measure, because they are distinct from military cooperation.

³The military contact variable on which the policy coordination score is based gives alliances that require wartime military contact a score of one, scores treaties with peacetime military contact as a two, and assigns alliances with defense cooperation a score of three. There is an order to the extent of policy coordination required as this variable increases.

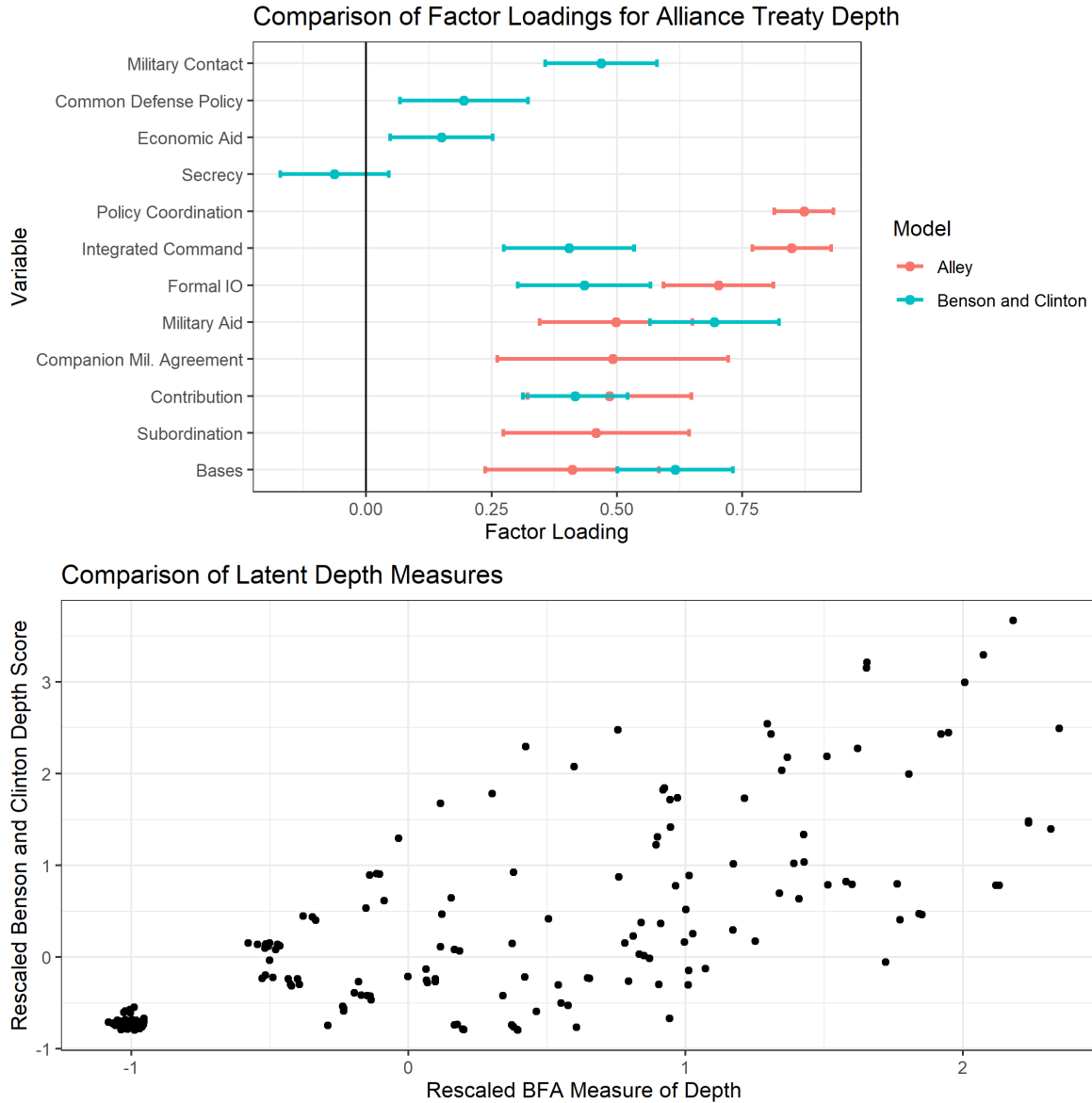


Figure 5: Comparison of latent measures of alliance treaty depth, one from this paper, and the other from Benson and Clinton (2016). The top panel compares the factor loadings from the two variables. The scatter plot compares latent depth scores from the semiparametric factor analysis with depth scores from Benson and Clinton (2016). The two latent measures have been rescaled by two standard deviations to facilitate comparisons. This comparison only includes alliances from version 3 of the ATOP data, because these are the alliances for which both models have scores.

Though the two measures of depth are correlated, they capture different concepts and my measure has fewer extreme outliers. Benson and Clinton address the general cost of an alliance, while I am focused on military coordination and cooperation. Benson and Clinton's measure is useful, but it departs from my aims in this project by including other variables and analyzing neutrality pacts. Because our measures operationalize different concepts in different groups of alliances, I believe my measure of depth is better suited for an analysis of the consequences of deep military cooperation in defensive and offensive alliances. My measure is not generally superior, and which latent depth measure scholars use in other analyses should depend on how they conceptualize alliance treaty depth.

References

- Benson, Brett V and Joshua D Clinton. 2016. "Assessing the Variation of Formal Military Alliances." *Journal of Conflict Resolution* 60(5):866–898.
- DiGiuseppe, Matthew and Paul Poast. 2016. "Arms versus Democratic Allies." *British Journal of Political Science* pp. 1–23.
- Juárez, Miguel A and Mark FJ Steel. 2010. "Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions." *Journal of Business & Economic Statistics* 28(1):52–66.
- Leeds, Brett Ashley and Sezi Anac. 2005. "Alliance Institutionalization and Alliance Performance." *International Interactions* 31(3):183–202.
- Murray, Jared S, David B Dunson, Lawrence Carin and Joseph E Lucas. 2013. "Bayesian Gaussian Copula Factor Models for Mixed Data." *Journal of the American Statistical Association* 108(502):656–665.