Appendix: Reassessing the Public Goods Theory of Alliances

December 9, 2020

This appendix contains supporting materials for the test of Hypotheses 1 and 2 in "Reassessing the Public Goods Theory of Alliances." Section 1 provides a demonstration that analyses of GDP and defense burdens are misspecified. Section 2 then provides more detail about the variables in the model and Section 3 assesses model fit and accuracy. The final section summarizes a test that uses average economic weight in their alliances to predict percentage changes in military spending.

1 Specification of GDP and Defense Burdens

As I note in the introduction, models and correlations between GDP and defense burdens are misspecified. Because defense burdens include GDP in the denominator, changes in GDP affect the expected value of defense burdens. When GDP shifts, the defense burden remains constant only if military spending also changes in such a way that defense spending's share of GDP remains the same. Such changes are highly unlikely.

The deterministic component in the relationship between GDP and defense burdens has important consequences for correlation and linear regression estimates. For a correlation ρ between two variables X and Y:

$$\rho = \frac{Cov(X, Y)}{Var(X)Var(Y)} \tag{1}$$

In a linear regression with one independent variable, the coefficient β_1 is equal to:

$$\beta_1 = \frac{Cov(X, y)}{Var(X)} \tag{2}$$

Correlations and regression coefficients depend on the covariance between the two variables. In general, covariance is equal to:

$$Cov(X, Y) = E[X, Y] - E[X]E[Y] \tag{3}$$

Now, consider the covariance between GDP and military spending as a share of GDP, $\frac{ME}{GDP}$.

$$Cov\left(GDP, \frac{ME}{GDP}\right) = E\left[GDP, \frac{ME}{GDP}\right] - E[GDP]E\left[\frac{ME}{GDP}\right] \tag{4}$$

The approximate expected value of the defense burden[1] is equal to:

$$E\left[\frac{ME}{GDP}\right] = \frac{E[GDP]}{E\left[\frac{ME}{GDP}\right]} - \frac{Cov(GDP, ME)}{\left(E\left[\frac{ME}{GDP}\right]\right)^2} - \frac{Var(ME)E[GDP]}{\left(E\left[\frac{ME}{GDP}\right]\right)^3} \tag{5}$$

Substituting this expectation into the covariance of GDP and defense burdens:

$$\begin{aligned} Cov\left(GDP, \frac{ME}{GDP}\right) = E\left[GDP, \frac{ME}{GDP}\right] - \\ E[GDP]\left(\frac{E[GDP]}{E\left[\frac{ME}{GDP}\right]} - \frac{Cov(GDP, ME)}{\left(E\left[\frac{ME}{GDP}\right]\right)^2} - \frac{Var(ME)E[GDP]}{\left(E\left[\frac{ME}{GDP}\right]\right)^3}\right) \end{aligned} \tag{6}$$

[1] Based on: `http://www.stat.cmu.edu/~hseltman/files/ratio.pdf`.

This then simplifies to:

$$Cov\left(GDP, \frac{ME}{GDP}\right) = E\left[GDP, \frac{ME}{GDP}\right] - \frac{(E[GDP])^2}{E\left[\frac{ME}{GDP}\right]} - \frac{E[GDP]Cov(GDP, ME)}{\left(E\left[\frac{ME}{GDP}\right]\right)^2} - \frac{Var(ME)(E[GDP])^2}{\left(E\left[\frac{ME}{GDP}\right]\right)^3} \quad (7)$$

Therefore the expected value of GDP, $E[GDP]$, affects the expected value of defense burdens $E\left[\frac{ME}{GDP}\right]$. This change in the expected value of defense burdens impacts the covariance between GDP and defense burdens, which then informs regression and correlation estimates. Because defense burdens are a non-linear function of military spending and GDP, the consequences of this specification problem for estimates are hard to predict, especially when GDP and military spending are correlated.

The implications of using a ratio variable like defense burdens for inference depend on how GDP and defense spending are correlated. At the very least, the GDP and defense spending values that Olson and Zeckhauser, as well as other researchers e.g.(Oneal, 1990; Kim and Sandler, 2019) employ in cross-sectional correlations depend on past GDP and military spending values. Temporal autocorrelation could generate correlations between GDP and military spending within units. If GDP and military spending decisions are correlated for other reasons, this further complicates the ratio calculations.

In what follows, I report the results of three simulation analyses to assess how cross-sectional correlations between GDP and defense burdens might compare to correlations between GDP and military spending. This simulation is a rough approximation of common tests of burden-sharing within NATO, which estimate annual correlations between GDP and defense burdens among NATO members. In all three simulations, I simulate an outcome and independent variable, both of which have strong temporal autocorrelation. I then analyze cross-sectional associations between the independent variable and the outcome with OLS. I compare inferences from measuring the outcome as a share of the independent variable to inferences without the ratio. I set the values of the

outcome and independent variable to roughly match the values of military spending burdens. The data-generating process contains 100 temporal observations of 20 units. Therefore, I run 100 cross-sectional analyses and each linear regression estimate has 18 degrees of freedom after estimating a regression coefficient and intercept.

In the first simulation, I assume that the independent variable and outcome are entirely uncorrelated, except for spurious associations from temporal autocorrelation. In the second simulation, I assume that the independent variable and outcome are correlated and each variable has temporal autocorrelation. I set the assumed correlation in the second simulation equal to the observed correlation between GDP and military spending in my data, which is .48. Because these analyses follow existing research designs by using cross-sectional snapshots, I do not control for temporal autocorrelation within units. The third simulation uses the simulated data from the second simulation, but converts the outcome into percentage changes, which is the outcome I use in the paper.

Inferences about the association between the simulated independent variable and its ratio with a dependent variable depend on whether the two variables are correlated, as Figure 1 shows. This figure plots whether the coefficient in a regression of simulated ratio and non-ratio outcomes on a simulated independent variable is statistically significant at conventional levels. When the independent variable and outcome are independent, creating a ratio generates an overwhelming number of statistically significant findings. When the independent variable and dependent variable are correlated, the ratio variable has a different effect. Creating a ratio of two correlated variables reduces the probability of a statistically significant result, relative to non-ratio estimates. Last, comparing the percentage change measure to a ratio outcome, the ratio outcome is more likely to produce statistically significant regression coefficients. Based on a chi-squared test, there are statistically significant differences in inferences from ratio and non-ratio outcomes for both types of simulated variables.

In summary, creating a ratio of the dependent and independent variable is likely to produce spurious results. If the two variables are uncorrelated, then estimates may conclude the two variables

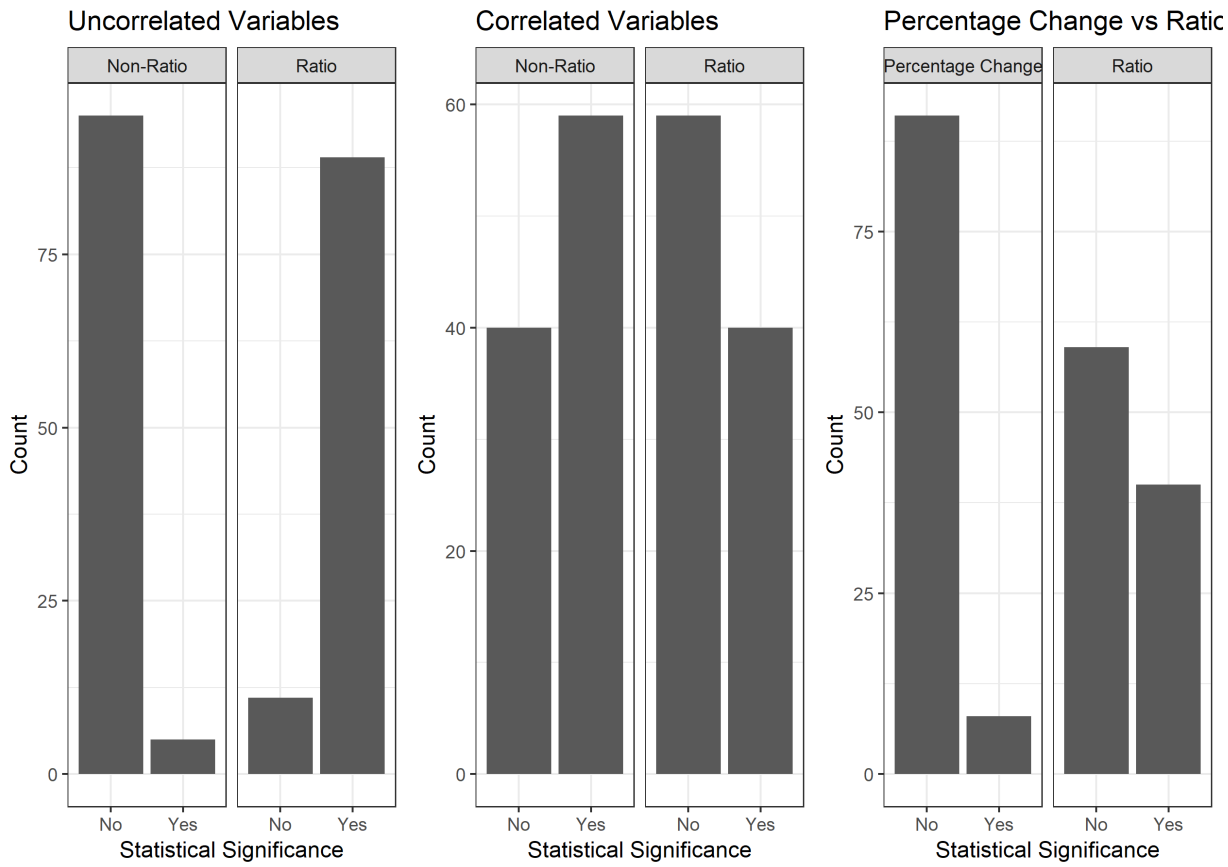

Figure 1: Estimated statistical significance of a linear regression coefficient in simulated cross-sectional analyses of ratio and non-ratio variables.

are correlated, when any association is driven entirely by changes in the independent variable. But if the two variables are correlated, ratio outcomes are more likely to produce findings that do not meet conventional statistical significance thresholds. Last, ratio variables are more likely to produce statistically significant estimates than the percentage changes variable this manuscript and Plümper and Neumayer (2015) prefer. Therefore, estimates of the correlation between GDP and military spending as a share of GDP may produce misleading findings, including null results when GDP and military spending are correlated.

2 Variables

Olson and Zeckhauser use GDP to measure state size, so I constructed a measure of GDP using data from the Maddison Project, which provides longer historical coverage (Bolt et al., 2018). I use military spending data from the Correlates of War Project (Singer, 1988). All alliance membership data comes from Version 4 of the Alliance Treaty Obligations and Provisions (ATOP) data (Leeds et al., 2002).

The dependent variable is percentage changes in military spending. Olson and Zeckhauser use defense spending as a share of GDP as their dependent variable, which is the source of previously described model specification problems (Plümper and Neumayer, 2015). I use percentage changes instead of the defense burden because this measure gives a sense of burdens from changing defense budgets, but has a lower risk of spurious inferences. Annual percentage changes in spending is the change in military spending as a share of the previous year's budget:

$$\% \text{ Change Military Spending} = \frac{\text{Mil. Ex.}_t - \text{Mil. Ex.}_{t-1}}{\text{Mil. Ex.}_{t-1}} = \frac{\Delta\text{Mil. Ex.}}{\text{Mil. Ex.}_{t-1}} \tag{8}$$

Measuring percentage changes in spending matches Olson and Zeckhauser's emphasis on how alliance participants allocate resources to the military. Positive percentage changes in spending imply an expanding defense budget and higher defense burden, all else equal. Moreover, using

percentage changes in spending mitigates the risk of spurious inferences due to non-stationarity in panel data. The log-level of military spending is not mean-reverting in long panels. A differenced military spending variable has increasing variance over time, as budgets expand and generate larger changes. Modeling the DV in levels or changes might lead to spurious inferences (Granger and Newbold, 1974).

Using percentage changes in military spending as the dependent variable benchmarks changes to budget size. This facilitates comparisons across states and years. A 2% change is an equally burdensome increase in the defense budget for large and small states, all else equal.

Besides the economic weight values in **Z**, I controlled for other variables that are correlated with alliance participation and military spending. I adjusted for international war (Reiter, Stam and Horowitz, 2016), civil war participation (Sarkees and Wayman, 2010), and a count of annual MIDs (Gibler, Miller and Little, 2016). I also included measures of regime type, external threat (Leeds and Savun, 2007), GDP, and the Cold War era.

3 Multilevel Model

This section describes the priors on the multilevel model, convergence diagnostics for the Hamiltonian Monte Carlo, and results from running the same model with a weighted economic size in the alliance participation matrix.

3.1 Priors

All priors are specified to be weakly informative relative to the scale of the data (Gelman, Simpson and Betancourt, 2017). I summarize the prior distributions for each set of parameters in Table 1. $p(\nu)$ is a well-behaved prior for the degrees of freedom in a t-distribution (Juárez and Steel, 2010).

Given that the median percentage change in military expenditures is 0.06, these normal priors

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

Table 1: Summary of Priors in Multilevel Model

are weakly informative, which means that they assume both small and extremely large effects are possible. I use normal priors because they cover the full range of likely parameter values and are easier to fit than heavier-tailed distributions.

To facilitate estimation, I use a non-centered parameterization for the state and year varying intercepts, as well as the γ parameters (Betancourt and Girolami, 2015). A non-centered parameterization decouples the mean and variance to express an equivalent prior, which makes sampling easier. I also employ a sparse matrix representation of the alliance membership matrix $\mathbf{Z}$ to speed up estimation.

3.2 Convergence

There were no divergent iterations in sampling. However, there are other threats to inference from the posterior samples. Given heavy tails in percentage changes of military spending, STAN might have struggled to explore the posterior distribution.

Energy plots can diagnose this problem. Figure 2 plots the marginal energy distribution and the first differenced distribution. If the two histograms do not overlap, sampling was impeded by heavy tails. The substantial overlap in the distributions for all four chains in Figure 2 indicates this was not a problem.

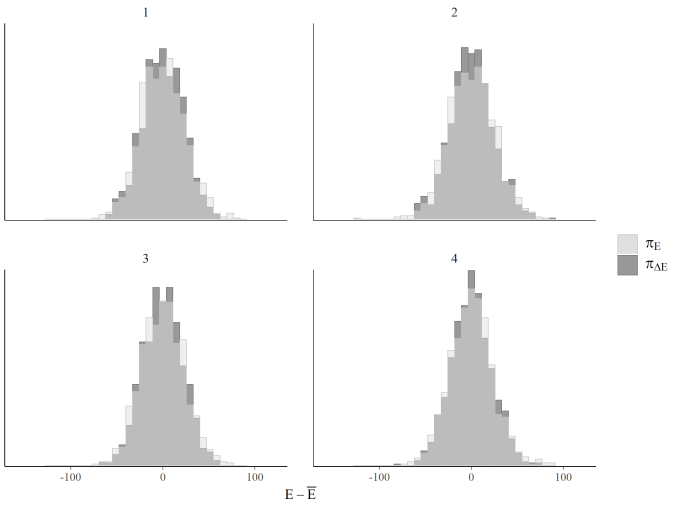

Figure 2: Energy plot of multilevel model results. Greater overlap in the two histograms indicates adequate exploration of the posterior distribution.

The split $\hat{R}$ statistic is another way to assess convergence. $\hat{R}$ compares the behavior of each chain by measuring the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains. When $\hat{R}$ is close to 1, all the chains have similar variance, and are therefore in equilibrium.

The standard heuristic is that an $\hat{R}$ greater than 1.1 is problematic. Figure 3 plots the $\hat{R}$ statistic for every parameter in the model. No parameters generate concern, even at a more conservative threshold of 1.05.

3.2.1 Inferences from Simulated Data

To assess if the model gives reasonable answers, I simulated data and associated parameters, then re-estimated the model on the simulated data. The model is a good fit if the credible intervals contain the known parameter values for the simulated data. This process checks whether the model can recover parameters from a known data-generating process that matches the model.

I simulate a hypothetical dataset with 2000 observations of 50 states observed over 200 years. I used part of the observed alliance data from the paper, due to problems simulating an alliance

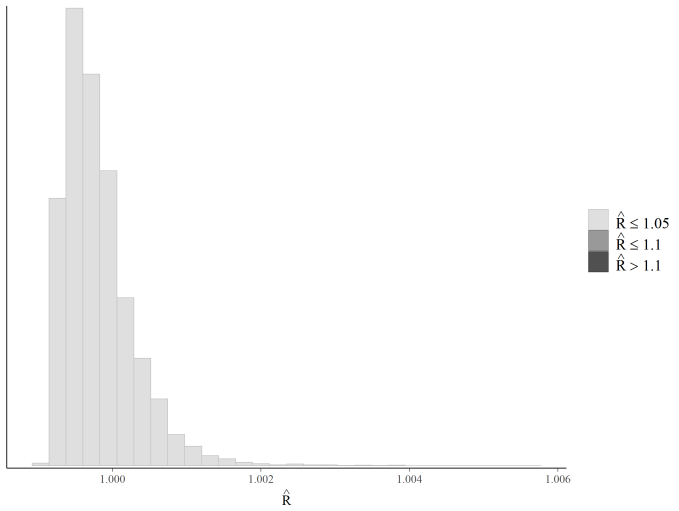

Figure 3: Histogram of split $\hat{R}$ statistic for all parameters in the multilevel model.

membership matrix with the same characteristics as the alliance data. There are 100 alliances in this data along with 2 state-level control variables. The hypothetical outcome is drawn from a Cauchy distribution with mean 0 and a scale of .25, which is more heavy-tailed than even my observed data.

I then simulate 2,000 draws of the outcome using the generated quantities block in STAN. The next step is selecting one of those draws of the outcome— which includes the value of the outcome for each observation and the associated parameter values. I select the 12th draw from the posterior and check whether after estimating the model on these data, the credible intervals include zero.

I focus on inferences about the γ, θ and σ_{all} parameters, because all three affect my test of the public goods argument. As Figure 4 and Figure 5 show, the posteriors accurately capture the known values of the hyper-parameters θ and σ_{all}. In these figures, the true parameter value is marked with a thick black line, while the light gray shaded area shows the 90% credible interval.

Because graphical presentation of the 100 γ parameters is more difficult, I calculated whether the credible interval contained the known parameter. 88 of the 100 intervals include the "true" γ value. Given the number of parameters and potential simulation variance, such accuracy is tolerable. Simulating data and recovering known parameters shows that the model estimates are

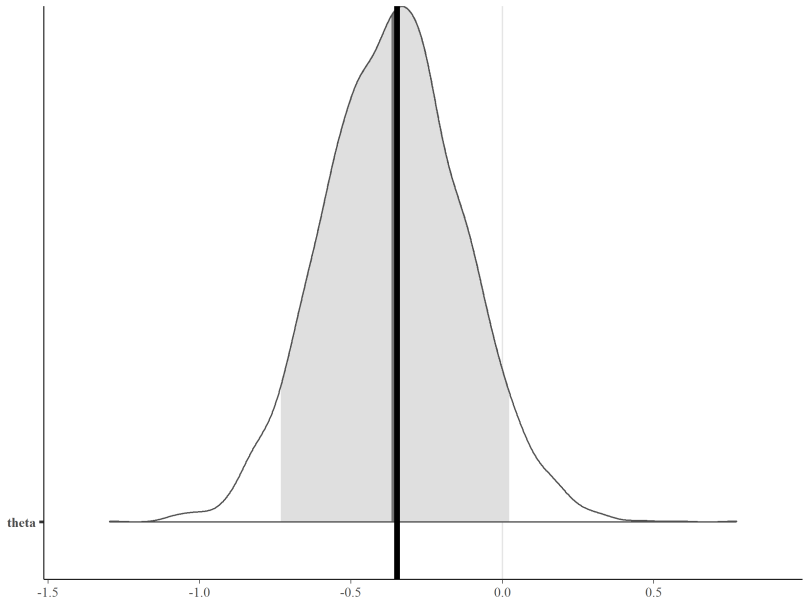

Figure 4: Posterior estimates and known parameter value for the alliance hyperparameter θ. The dark gray bar marks the posterior mean, while the shaded area captures the 90% credible interval. The black line marks the known, "true" θ value, which falls within the 90% interval.

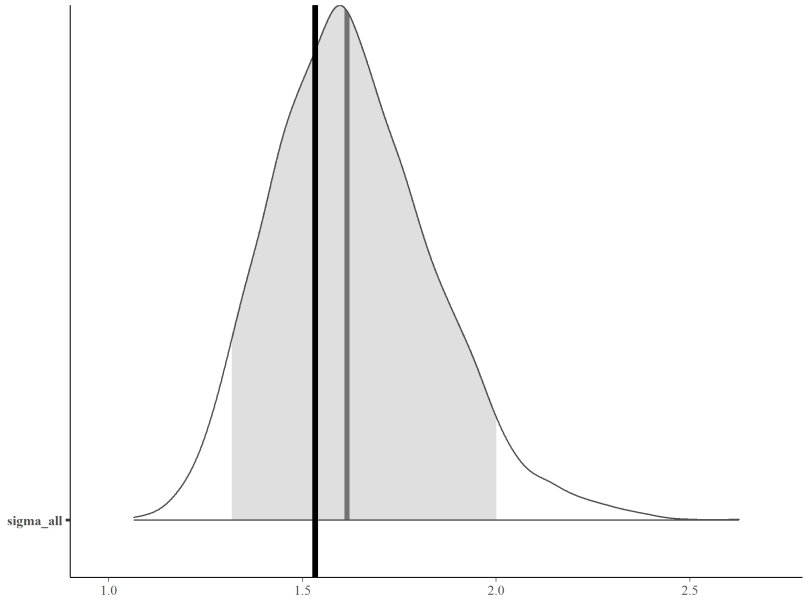

Figure 5: Posterior estimates and known parameter value for the alliance hyperparameter σ_{all}. The dark gray bar marks the posterior mean, while the shaded area captures the 90% credible interval. The black line marks the known, "true" σ_{all} value, which falls within the 90% interval.

reasonable approximations of the data-generating process.

3.3 Substantive Effect Calculations

To further examine whether increasing a state's share of allied GDP leads to higher defense spending, I simulated the effect of changing economic weight on percentage changes in military spending in the multilevel Bayesian model. In the simulated data, I used the full posteriors of the intercept α, all the β coefficients, and one γ parameter. I selected the economic weight parameter with the most positive posterior mass, so this is the *best case alliance* for the public goods models. I then set the state-level variables at their median or modal value and changed economic weight from -1 to 1.

In Figure 6, I summarize predicted changes in military spending at the two economic weight values. In this figure, the point marks the mean and the error bars summarize the 90% credible interval. There is limited evidence that larger alliance participants have higher military spending.

3.4 Alternative Coding of Economic Size

The values of -1 for small states and 1 for large states in alliance participation matrix **Z** in the manuscript results create coarse bins. This could mask differences within each category that affect inferences about economic weight and percentage changes in military spending within alliances. This section checks whether inferences are sensitive to an alternative coding of **Z**.

To retain the same split of negative values for small states and positive values for large states, I subtracted one from economic weights that fell below the median in bilateral or multilateral alliances. This means that states with a small share of allied GDP have larger negative values than states with close to the median value. For example, a state with a 25% of total allied GDP in a bilateral alliance has a weight of -.75 with this variable, but a state with 49% of allied GDP has a weight of -.51. Positive values like .51 are unchanged.

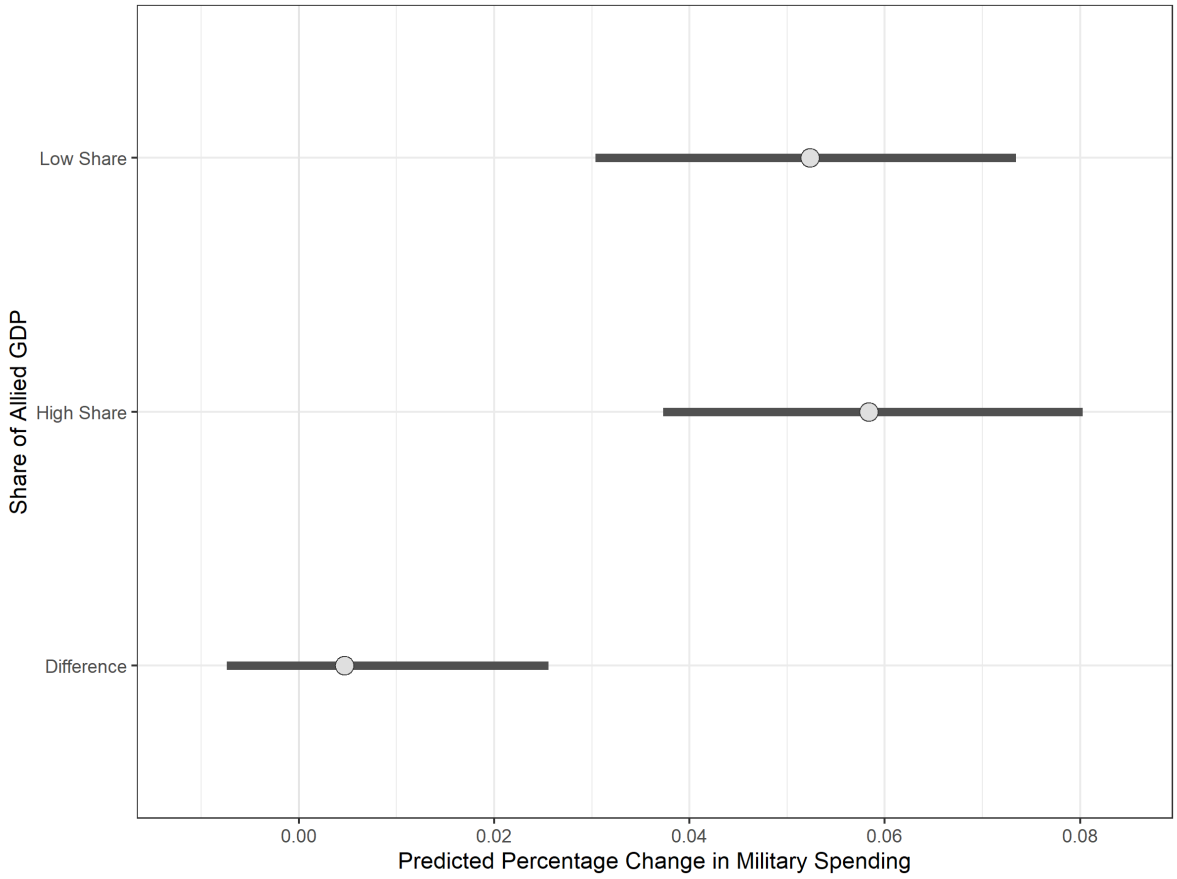

Figure 6: Predicted percentage changes in military spending for a simulated state with low or high shares of total allied GDP. The public goods model predicts a positive difference between the high and low economic weight scenarios. Points mark the median value, and the error bars summarize the 90% credible interval. The difference estimate captures the effect of moving from low to high economic weight.

Again, the public goods model would predict many positive γ parameters, as these would reflect higher military spending for large states and lower military spending for small states. As Figure 7 shows, there are few positive γ values. There are two alliances with a clearly positive γ parameter. These are the OAS (ATOPID 3150) and an alliance between the United States and Thailand (ATOPID 3260). 35 alliances have a positive posterior mean, as well.

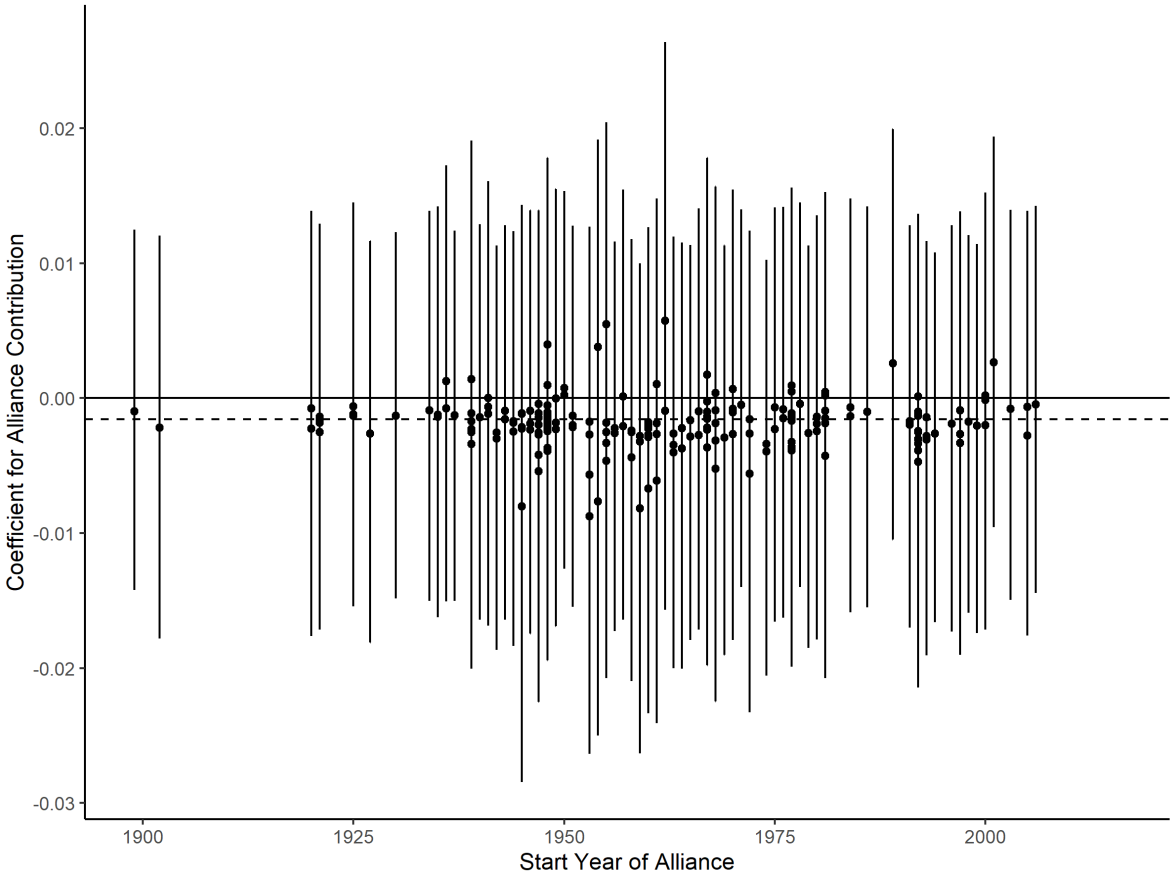

Figure 7: Estimated γ parameters from a model with negative and positive economic weights in the alliance membership matrix.

While this approach is somewhat better for the public goods model's predictions, the evidence is similar. There are very few alliances with even limited evidence that economic weight leads to higher military spending.

4 Average Economic Weight and Military Spending

The model I use in the paper estimates alliance-specific associations between economic weight and military spending. I also consider whether higher average economic weight across multiple alliances increases percentage changes in military spending. This section summarizes results from four models of the correlation between average economic weight and military spending. Following the recommendations of Rainey and Baissa (2020), I use OLS and robust regression estimators with both transformed and unaltered military spending growth. I use an inverse hyperbolic sine to transform the outcome because this transformation includes positive, negative and zero values. Robust regression down weights unusual observations, which is important because the residuals of OLS estimators in this data are not normally distributed.

The key independent variable is a state's average economic weight across all of its alliances. I also include measures of average alliance size and democracy (DiGiuseppe and Poast, 2016) as controls. Other controls are the same as the multilevel model.

Table 2 summarizes the results of these four estimation strategies. The robust regression coefficients for average economic weight are of small substantive magnitude and the 95% confidence intervals include a range of positive and negative values. The OLS estimates for average economic weight are larger, but the difference in residual standard error between these estimates and the robust regression suggests that the OLS estimates are driven by unusual observations.

To asses the substantive impact of increasing a state's average economic weight in its alliances, I simulated the effect of moving average weight from the first quartile (0.05) to the third quartile (.40). Holding all other variables at their medians or modes, this increase in weight implies an expected percentage change in military spending of -0.005. The 90% credible interval for this change ranges from -0.017 to 0.01, so the estimated impact of a massive rise in average economic weight includes large positive, large negative and null effects.

	Dependent variable:			
	% Change Milex.		IHS(% Change Milex.)	
	robust linear	*OLS*	*OLS*	*robust linear*
	(1)	(2)	(3)	(4)
Avg. Economic Weight	−0.008	0.382	0.030	−0.008
	(−0.051, 0.036)	(−1.024, 1.789)	(−0.058, 0.118)	(−0.051, 0.036)
ln(GDP)	−0.035	−6.656	−0.363	−0.033
	(−0.135, 0.066)	(−9.896, −3.417)	(−0.566, −0.160)	(−0.133, 0.066)
Avg. Alliance Size	−0.0001	−0.014	0.00004	−0.0001
	(−0.001, 0.001)	(−0.039, 0.011)	(−0.002, 0.002)	(−0.001, 0.001)
Avg. Allied Democracy	0.0003	−0.031	−0.001	0.0003
	(−0.001, 0.002)	(−0.084, 0.022)	(−0.005, 0.002)	(−0.001, 0.002)
International War	0.082	0.149	0.140	0.080
	(0.054, 0.109)	(−0.728, 1.026)	(0.085, 0.195)	(0.053, 0.107)
Civil War Participant	−0.003	0.178	0.020	−0.003
	(−0.023, 0.017)	(−0.464, 0.821)	(−0.020, 0.061)	(−0.023, 0.017)
Regime Type	−0.001	0.039	0.0002	−0.001
	(−0.002, 0.001)	(−0.002, 0.079)	(−0.002, 0.003)	(−0.002, 0.001)
External Threat	0.072	1.109	0.107	0.071
	(0.040, 0.104)	(0.071, 2.146)	(0.042, 0.172)	(0.039, 0.103)
Cold War	0.040	0.439	0.040	0.040
	(0.029, 0.052)	(0.058, 0.820)	(0.016, 0.064)	(0.028, 0.052)
Constant	0.067	5.420	0.353	0.065
	(−0.015, 0.148)	(2.791, 8.049)	(0.189, 0.518)	(−0.016, 0.146)
Observations	5,022	5,022	5,022	5,022
R^2		0.007	0.015	
Adjusted R^2		0.005	0.013	
Residual Std. Error (df = 5012)	0.160	6.230	0.390	0.159
F Statistic (df = 9; 5012)		3.657 (p = 0.0002)	8.219 (p = 0.000)	
Note:				95% Confidence Intervals in Parentheses.

Table 2: OLS and robust regression of the association between average economic weight in alliances and percentage changes in military expenditures from 1919 to 2007.

5 Small States Joining Existing Alliances

Another way to look at free-riding is to consider the results of small states joining an existing alliance.[2] In these cases, if small states are free-riding, they may be able to lower their defense spending. I consider two cases of small states joining an existing alliance— NATO expansion and new Arab League members. Figure 8 tracks military spending growth before and after alliance participation for new members of these two alliances.

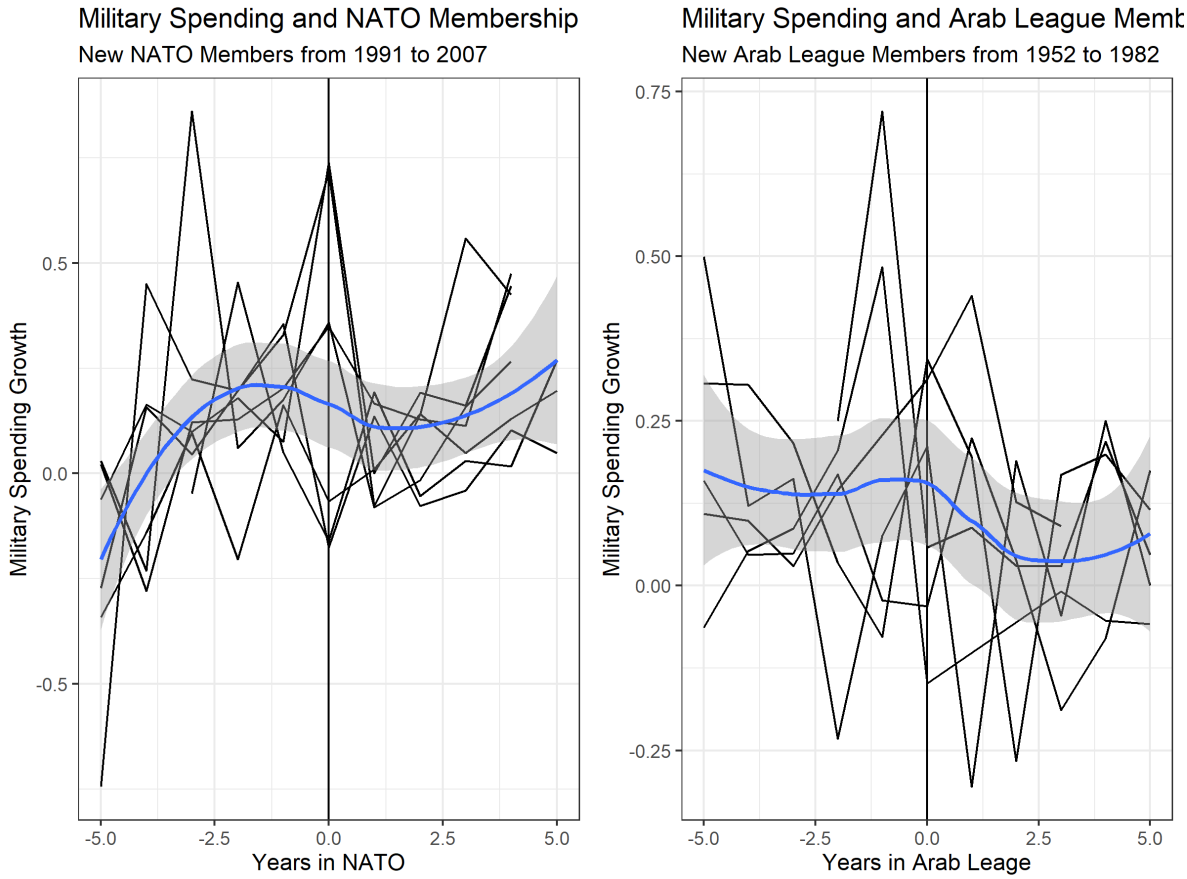

Figure 8: Percentage changes in military spending in the five years before and after joining NATO and the Arab League for states with a small share of allied GDP. Each line shows percentage changes in military spending by a state over time, and lines that stop before five years in the alliance reflect missing data or the end of the sample. The vertical line at year 0 marks when a state joined the alliance. A loess curve summarizes the overall trend across multiple states.

[2]Thanks to an anonymous reviewer for suggesting this analysis.

There is little evidence that small states have lower percentage changes in military spending after joining NATO. Before and after NATO membership, most of the European states that joined NATO had positive spending growth. There is not an obvious drop in annual defense allocations, even with substantial budget instability in the aftermath of the Cold War. Several of the Baltic states actually increased defense spending as part of the US-led coalition in Iraq.

The Arab League shows more evidence of falling defense budgets among new small members. Some junior members of the Arab League had either smaller increases or reductions in their defense budgets after joining the alliance. As in the case of new NATO members, these states had large defense budget instability.

References

Betancourt, Michael and Mark Girolami. 2015. Hamiltonian Monte Carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications*, ed. Satyanshu K. Upadhyay, Umesh Singh, Dipak K. Dey and Appaia Loganathan. Chapman and Hall/CRC Press pp. 79–102.

Bolt, J, R Inklaar, H de Jong and JL van Zanden. 2018. "Maddison Project Database, Version 2018." *Rebasing 'Maddison': new income comparisons and the shape of long-run economic development* .

DiGiuseppe, Matthew and Paul Poast. 2016. "Arms versus Democratic Allies." *British Journal of Political Science* pp. 1–23.

Gelman, Andrew, Daniel Simpson and Michael Betancourt. 2017. "The prior can generally only be understood in the context of the likelihood." *arXiv preprint arXiv:1708.07487* .

Gibler, Douglas M, Steven V Miller and Erin K Little. 2016. "An Analysis of the Militarized Interstate Dispute (MID) Dataset, 1816–2001." *International Studies Quarterly* 60(4):719–730.

Granger, Clive WJ and Paul Newbold. 1974. "Spurious Regressions in Econometrics." *Journal of Econometrics* 2(2):111–120.

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.

Kim, Wukki and Todd Sandler. 2019. "NATO at 70: Pledges, Free Riding, and Benefit-Burden Concordance." *Defence and Peace Economics* pp. 1–14.

Leeds, Brett Ashley and Burcu Savun. 2007. "Terminating Alliances: Why Do States Abrogate Agreements?" *The Journal of Politics* 69(4):1118–1132.

Leeds, Brett, Jeffrey Ritter, Sara Mitchell and Andrew Long. 2002. "Alliance Treaty Obligations and Provisions, 1815-1944." *International Interactions* 28(3):237–260.

Oneal, John R. 1990. "The theory of collective action and burden sharing in NATO." *International Organization* 44(3):379–402.

Plümper, Thomas and Eric Neumayer. 2015. "Free-riding in alliances: Testing an old theory with a new method." *Conflict Management and Peace Science* 32(3):247–268.

Rainey, Carlisle and Daniel K. Baissa. 2020. "When BLUE Is Not Best: Non-Normal Errors and the Linear Model." *Political Science Research & Methods* 8(1):136–148.

Reiter, Dan, Allan C. Stam and Michael C. Horowitz. 2016. "A Revised Look at Interstate Wars, 1816–2007." *Journal of Conflict Resolution* 60(5):956–976.

Sarkees, Meredith Reid and Frank Whelon Wayman. 2010. *Resort to War: A Data Guide to Inter-state, Extra-state, Intra-state, and Non-state Wars, 1816-2007*. Washington, DC: CQ Press.

Singer, J David. 1988. "Reconstructing the correlates of war dataset on material capabilities of states, 1816–1985." *International Interactions* 14(2):115–132.