Abstract

Berkemeier and Fuhrmann (2018) discover that only 22.2% of alliances fulfill their commitments post-World War II, far lower than the 74.5% compliance rate Leeds (2003) find for 1815-1944. Operationalization choices – not empirical or modeling assumptions – drive this difference. The alliance reliability literature is conditioned on "performance opportunities." Consequently, it misses two historical trends: Security pacts have increased in size over the past two centuries, and the incidence of war has declined. Correcting for this, reliability estimates vary widely, from just over 1% to nearly 90%. This research note assesses how these trends affect reliability estimates, and it examines a measure of alliance cohesion/fragmentation that can circumvent these issues, also drawn from the Alliance Treaty Obligations and Provisions dataset. Finally, it reevaluates Leeds and Anac (2005), contradicting its core finding that military institutionalization reduces alliance reliability.

1 Introduction

Leeds (2003) find that 74.5% of alliances from 1815-1944 honor their commitments. Berkemeier and Fuhrmann (2018) extend that analysis to 2003, discovering a much lower rate: Only half fulfill their promises. This is driven by a 22.2% post-World War II compliance rate. Yet, states created nearly twice as many alliances in that 59-year period than the 130 years prior. How can we explain the juxtaposition of high creation and low fulfillment?

This puzzle results from operationalization choices, not empirical or modeling as-

sumptions. The alliance reliability literature is conditioned on "performance opportunities." According to Sabrosky (1980), Leeds (2003), and Berkemeier and Fuhrmann (2018), only when a pact is invoked under the proper circumstances should we begin to assess whether states are keeping their promises, as detailed in Section 2. But this analytical condition misses two long-term trends. First, the risk of war has declined. From 1815-1944, the baseline chance an alliance would confront a performance opportunity was 41.75%. That fell to 9.47% post-World War 2. Most contemporary alliances have never been tested in war, and these "unchallenged" pacts are censored from the analysis and systematically differ from both honored and abrogated agreements. Consequently, reliability estimates may not generalize to different sets of alliances, states, and even time periods.

Second, alliances have grown larger across the two centuries covered in the Alliance Treaty Obligations and Provisions (ATOP) dataset. Relative to the size of the state system, a smaller number of pacts now encompass many more members. But by conditioning on an alliance-level concept like performance opportunity, the literature overweights bilateral pacts and underweights multinational ones. In 2003, nearly 90% of alliances were bilateral. But in that same year, nearly 90% of security *relationships* ("alliance-dyads") were multinational.

Reliability estimates are highly sensitive to these operationalization choices (i.e. whether to censor and what unit of observation to use). Section 3 details these historical trends and adjustments to these choices, demonstrating that alliance fulfillment can range from just over 1% to nearly 90%. Section 4 presents alternative measures of alliance cohesion/fragmentation – *Term* and *Term Cause*, also drawn from ATOP – that elide these problems. They can also explain why states created so many post-

World War 2 pacts despite Berkemeier and Fuhrmann (2018)'s low reliability estimate. With a declining incidence of war, states worried relatively less about demonstrated reliability towards external threats and relatively more about internal cohesion. And fragmentation and collapse is infrequent, ranging from 16.9% to 1.14%.

Section 5 replicates and reevaluates Leeds and Anac (2005) in light of these findings. That study argued that military institutionalization and formalization reduced alliance fulfillment, contradicting the costly signaling and rational institutionalism approaches.[1] However, institutionalization in particular is theoretically connected to both alliance size (e.g. Wallander and Keohane (1999)) and reliability (e.g. Fearon (1997)). If Leeds and Anac (2005) are correct, then changing the unit of observation from "alliance" to "alliance-dyad" should increase the proportion of both alliance institutionalization *and* violation in the dataset. Instead, this change produces the opposite effect: Institutionalization improves the likelihood of states honoring their promises.

The conclusion summarizes the general implications of the two historical trends for the alliance reliability literature, as well as any studies that draw upon military pacts as dependent or independent variables.

2 Alliance Reliability

Sabrosky (1980) and Siverson and King (1980) initiated the alliance reliability literature. Critically, they and the rest of this literature conditioned their calculations

[1] Fearon (1997); Morrow (2000); Koremenos, Lipson and Snidal (2001); Koremenos (2016).

on the "alliance war performance opportunity." Reliability can only be determined if and when an alliance is confronted with war. They reach the pessimistic conclusions that military pacts are honored only 27% and 23.1% of the time, respectively.

Leeds, Long and Mitchell (2000) note that Sabrosky (1980) in particular assumes that "every alliance requires the partners to fight together regardless of the context of the war."[2] This would not apply, for example, to consultative and neutrality pacts, which do not necessarily require allies to fight on behalf of one another. The authors code the *casus foederis* for military commitments in the Alliance Treaty Obligations and Provisions dataset, calculating a substantially higher 74.5% fulfillment rate for alliances from 1815–1944.[3]

Finally, Berkemeier and Fuhrmann (2018) extend the Leeds, Long and Mitchell (2000) analysis from 1945–2003. Since World War 2, only 22.22% of alliances are honored.[4] The authors suggest a few reasons for this precipitous drop in reliability. Offense and neutrality pacts are honored at higher rates than defensive ones, and those pacts are much rarer post-1945. But the fulfillment of defensive agreements also fell, from 61.02% (1816-1944) to 13.95% (1945-2003). They conjecture that, because of nuclear weapons, states rarely challenge great powers and their defensive commitments. Consequently, performance opportunities disproportionately occur against weak states, who often lack the material capability to decisively sway war outcomes. So they defect.

ATOP sparked many recent studies on the determinants of alliance reliability. In

[2] Leeds, Long and Mitchell (2000, 691).
[3] Leeds et al. (2002).
[4] They also calculate a slightly lower fulfillment rate from 1815-1944 than do Leeds, Long and Mitchell (2000), 66.3% compared to 74.5%.

contrast to the rational design and costly signaling literatures, Leeds and Anac (2005) finds that formalization and military institutions reduce the likelihood of alliance fulfillment.[5] Mattes (2012), however, contends that states employ institutionalization strategically. Countries with a history of prior alliance defections use organizational sunk costs to demonstrate future reliability, particularly within "symmetric" alliances comprised of countries with similar power levels. She demonstrates this using statistical analysis of bilateral alliances between 1919 and 2001.

States also design alliances so they can more easily be fulfilled. Benson (2012) notes that states often include conditionality provisions limiting when a pact can be invoked (the *casus foederis*), while Kim (2011) highlights the similar function of ambiguous language. These same concerns disproportionately drive democracies to adopt consultative provisions to avoid costs of violation.[6]

The (lack of) reliability affects a wide range of interstate security behavior. According to Leeds (2003), alliances deter conflict by revealing the scope of retaliation adversaries can expect. Johnson and Leeds (2011) claim that defense pacts are particularly effective at deterring conflict initiation without simultaneously inducing moral hazard by partners. But low alliance fulfillment rates – as calculated by Berkemeier and Fuhrmann (2018), Sabrosky (1980) and Siverson and King (1980) – call into question the validity of these findings.

[5] Koremenos, Lipson and Snidal (2001); Koremenos (2016); Fearon (1997); Morrow (2000).
[6] Chiba, Johnson and Leeds (2015).

3 Data Censoring and Alliance Size

These low fulfillment rates present an even broader puzzle. States established 215 alliances in the 130 years between 1815-1944, or approximately 33.2% of the observations in the ATOP 3.0 dataset. By contrast, they established 433 pacts in the 59 years between 1945-2003. Why did they create so many more partnerships in that shorter time span if they expected lower fulfillment rates?

Rather than a theoretical or an empirical puzzle, this article suggests the answer lies in operationalization choices. In predicating its analysis on performance opportunities, the reliability literature is misaligned with two long-term historical trends: the declining incidence of interstate war and increasing alliance size. Depending on how scholars structure their data, these trends can produce wildly varying estimates of alliance reliability. The calculations below range from fulfillment rates of just over 1% to nearly 90%.

The first historical trend is the declining incidence of war. As mentioned, the literature is conditioned on "alliance war performance opportunities," an alliance-level variable. From 1815-2003, Berkemeier and Fuhrmann (2018) identify 576 instances where an alliance's *casus foederis* was invoked due to war. But these opportunities are not distributed evenly across time, and interstate war has declined significantly since World War 2. Out of 454 post-1945 alliances, only 43 had a performance opportunity. By contrast, 81 out of 194 alliances invoked their *casus foederis* prior to that year. The "base rate" of a military pact experiencing at least one performance opportunity fell from 41.75% to 9.47%. Consequently, the majority of post-World War 2 alliances

are censored from reliability calculations. Indeed, citing Schelling (1966), Berkemeier and Fuhrmann suggest that censored pacts may be particularly good at deterrence. "The most effective threat is one that never has to be implemented. NATO does not appear in our dataset, for example, precisely because potential adversaries perceive it as effective." Nevertheless, they conclude that, when war occurs, states are less likely to honor their commitments than Leeds, Long and Mitchell (2000)'s earlier calculation anticipates.

But there are systematic differences between "un-challenged" alliances; challenged alliances whose members do not honor their commitments; and challenged, honored alliances. Table I visualizes this using the ATOP3.0 dataset. On average, unchallenged pacts (Column 1) have more democratic members engaged in significantly more trade, both of which are associated with decreased internal and external conflict.[7] These states also appear to be weaker than those in "challenged" ones, regardless of whether the latter violate or honor their commitments (Columns 2 and 3, respectively). Unchallenged alliances have a longer duration, their members participate in more alliances, and are larger than challenged, honored pacts, a point discussed further detail below. Each of these characteristics is theorized to have differential effects on deterrence and reliability.

We also see differences between "not honored" and "honored" alliances. The former tend to be weaker than the latter (in line with Berkemeier and Fuhrmann's suggestion), as well as more democratic. The pacts last longer, and, like unchallenged alliances, they are larger than honored ones. Consequently, these categories have significant differences of theoretical importance, such that estimates of reliability

[7] Gartzke (2007); Maoz and Russett (1993); Owen (1994).

Table I: Comparison of Unchallenged, Violated, and Honored Alliances. Average values presented.

	Unchallenged	Challenged, Not Honored	Challenged, Honored
Duration (Years)	27.78	22.77	13.86
Size	5.43	7.37	2.29
Number of Alliances	11	9	6
CINC Score	0.0104	0.0617	0.1101
Polity2	6	6	1
National Trade Flow	18188.99	4905.205	4288.285

based on challenged pacts may not be generalizable to unchallenged ones.

The second historical trend is that alliances have gotten larger. Consider Figure 1, which displays the proportion of alliances that are bilateral. At first glance, it appears as if bilateralism is the dominant feature of international security cooperation. In 2003, well over 80 percent of alliances had only two members, and Mattes (2012), Leeds and Savun (2007), Leeds, Mattes and Vogel (2009), and Kim (2011) develop their theories based on these bilateral pacts and security ties.

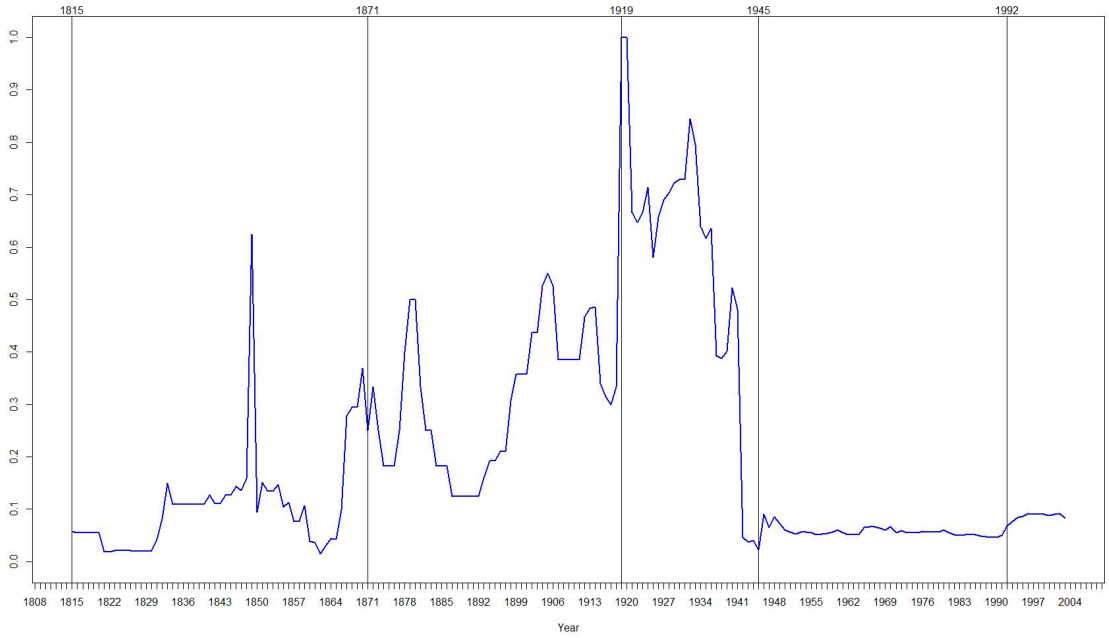
But Figure 1 is misleading if we are interested in alliances as a proxy for interstate security relations, rather than in alliances as individual institutions. It implicitly treats the bilateral U.S.-Japan and China-DPRK pacts as equivalent to the OAS, the Arab League, and NATO. Despite the latter encompassing far more states, all would have the same number of observations (i.e. one) in any dataset using "alliance" as the unit of analysis, as conditioning on performance opportunities implicitly does.

Figure 1: Frequency of Bilateral Alliances, 1815–2003. Alliance as unit of analysis.

Using *alliance-dyads* better reflects interstate security relations, as partially explored by Fordham and Poast (2014) and Poast (2010). That is, we break down each pact into its constituent interstate links, modeling the web of military relationships between states. The U.S.-Japan alliance would remain a single observation. But NATO, with its 28 members in 2003, has 378 observations (e.g. U.S.-UK, U.S.-Germany, UK-Germany, etc.), reflecting its significantly larger security network. Figure 2 converts the previous image into alliance-dyads. The impression is starkly reversed. While most *alliances* are bilateral, most *allies* engage in multilateral pacts. In 2003 (ATOP3.0's final year), 89.64 percent of alliances (173/193) were bilateral, but only 11.87 percent of alliance-dyads (177/1499) participated in such pacts. This makes intuitive sense. Given a fixed number of states, as alliances get larger, fewer pacts would encompass more members. Put another way, as pacts get bigger, there are

fewer "leftover" countries needing or able to create additional partnerships. But this consequently increases the proportion of bilateral treaties as a share of all *alliances*, as multinational pacts count as only a single body under that unit of observation.

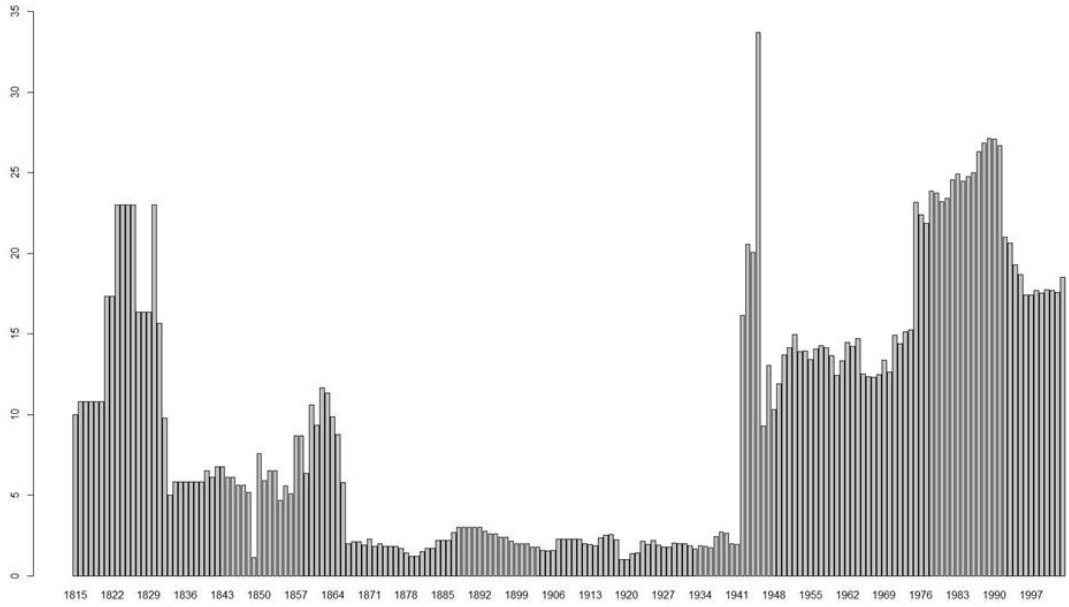
Figure 2: Frequency of Bilateral Alliances, 1815–2003. Alliance-dyads as unit of analysis.



Finally on this point, Figure 3 plots the average number of alliance members from 1815-2003. There is little analytical difference between using alliances or alliance-dyads if we restrict our sample to 1866–1939. Partnerships in these years had roughly 2-3 members. Outside of that range, however, alliances had a dozen members on average. From the 1970s onwards, many had twice that number. Moreover, scholarship demonstrates that size leads to systematic differences in institutions. Smaller alliances offer flexibility and even encourage defection, particularly as the direct reputational

damage from abrogation is limited when compared to multinational pacts.[8] Larger ones are more likely to use formal structures to coordinate members and reduce policy externalities.[9] They also aggregate greater power against a common threat, enhancing deterrence.[10]

Figure 3: Average alliance size by year, 1815–2003.



In sum, these two historical trends imply that fewer alliances are being challenged, and those alliances encompass many more states. To make comparisons across this time span, scholars must select a unit of analysis for their reliability calculations. But these choices can be more or less aligned with these historical trends, as Table II illustrates. Columns are divided into those using alliances and those using alliance-dyads as the unit of analysis. Rows include or exclude censored observa-

[8] Morrow (2000).
[9] Wallander and Keohane (1999); Koremenos, Lipson and Snidal (2001).
[10] Leeds (2003).

tions. Honor/violation conversions follow Leeds (2003), Leeds and Anac (2005), and Berkemeier and Fuhrmann (2018). If a single dyad reneges on its commitment, all of that alliance's dyads are coded as violated.[11] Censoring follows Berkemeier and Fuhrmann (2018): 272 ATOP alliances and 748 alliance-dyads are not included in their data, and Row 1 of Table II excludes these observations as well. In Row 2, censored observations are treated as honored, following Schelling (1966) and Berkemeier and Fuhrmann (2018).

Table II: Alliance Reliability Calculations

	All Observations		Post-World War 2	
	Alliance	Alliance-Dyads	Alliance	Alliance-Dyads
Honor, Data Censored	0.496 69/139	0.056 107/1906	0.208 11/53	0.014 12/876
Honor, Data Not Censored	0.878 506/576	0.506 1842/3641	0.887 330/372	0.64 1537/2401

Including censored data increases reliability rates: Within columns, "Honor, Data Censored" rates are always higher than "Honor, Data Not Censored." By contrast, adjusting for alliance-dyads reduces reliability estimates, in part because the literature counts any member's violation against the entire pact. Moreover, we see high variability in fulfillment rates. Using the entire dataset, these range from 5.61% (alliance-dyads with censoring) to 87.75% (alliances without censoring). Using post-World War 2 observations only, the range is 1.37% to 88.71%.

Overall, conditioning alliance reliability estimates on performance opportunities creates two broad analytical problems. First, it forces a trade-off between conceptual fidelity and generalizability. None of the operationalizations in Table II is the

[11] This tilts the analysis towards higher violation rates.

"right" one. Instead, scholars must align their research questions with the problem set that states define and face. Do diplomats regularly expect war, such that wartime reliability is a critical concern?[12] Or do they expect warfare to be unlikely, conditioning their reliability assessments on that probability? Are states concerned whether all their partners show up, which might be particularly important in smaller pacts? Or do they view alliances are as an integral component of broader interstate security relations, making certain military ties contingent on and/or less important than others?[13]

The answers to these questions will fit the data better for particular years, alliances, and operationalizations than for others. Certain conceptualizations and theories may demand specific operationalizations, but findings based on these decisions cannot be easily generalized to other pacts. Indeed, this is the motivating issue with Berkemeier and Fuhrmann (2018). The low post-1945 fulfillment rate is at odds with the high number of post-1945 alliances and their increasing average size. States appear to consider pacts much more reliable than many of Table II's calculations suggest.

This leads to the second analytical problem. As hinted above, our definition of reliability biases results against multilateral pacts being honored. The literature codes reliability as an alliance-level variable occurring "only when *all* members fulfill their obligations."[14] Defection by a single partner, no matter how peripheral, is treated as a violation. Because they have more members, multinational alliances likely experience a higher rate of defection simply due to coding strategy. Indeed, using Berkemeier and

[12] Schroeder (1976) and Lascurettes (2020) note that warfare was a regular and even legitimate state activity during the 17th and 18th centuries.
[13] Fordham and Poast (2014); Poast (2010).
[14] Leeds and Anac (2005, 193). Emphasis added.

Fuhrmann (2018)'s data, alliance fulfillment is positively correlated with "bilateralness" at 0.278. Alliances honoring their agreements have 2.29 members on average, while those violating them are over three times as large, with 7.37 members typically, as seen in Table I.

Treating violation as an alliance-level variable ignores that defections and fulfillment by certain members are more important than others.[15] Great power commitments, for example, should be more important to alliance outcomes. Leeds, Long and Mitchell (2000) and Berkemeier and Fuhrmann (2018) both count ATOP2550, or the Allies in World War 2, as violated. It is not clear which states abrogated their commitments, why, or when. The treaty was signed in 1942, so signatories knew they would be fighting the Axis powers. Moreover, 15 countries acceded after initial ratification. Despite these defections (and additions), the Allies accomplished their central goal of defeating the Axis powers. Counting the entire alliance as violated despite the manifest success of some members may not accurately reflect the underlying interstate nature of abrogation nor its effects.

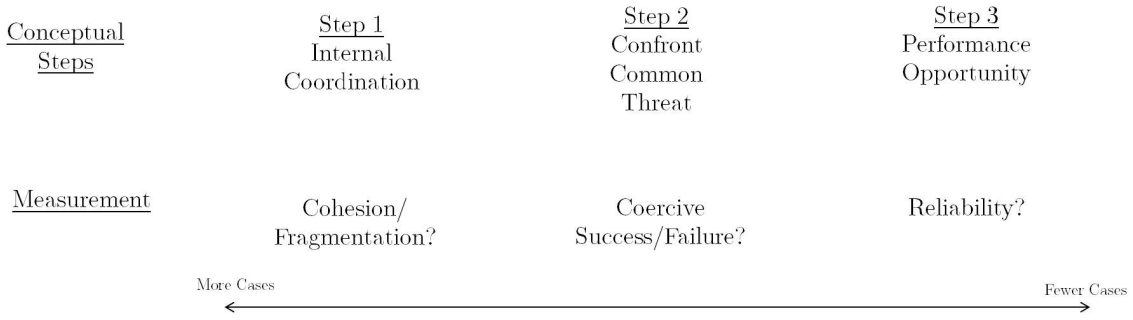
4 An Alternative to Reliability: Operationalizing Internal Alliance Cohesion

Yet, the literature has strong conceptual reasons to condition reliability on performance opportunities. It casts alliances as tools of external threat management activated when adversaries launch a military challenge. Performance opportunities result

[15] Poast (2010).

from cumulative selection processes visualized in Figure 4. The reliability literature concentrates on Step 3. Alliances have first resolved any internal coordination problems inhibiting pact formation and continuation (Step 1). Some of these partnerships may successfully coerce or deter their adversaries, such that they are not challenged (Step 2). Some, however, may fail, generating a performance opportunity. The literature finally evaluates their reliability in Step 3.

Figure 4: Cumulative Alliance Selection Processes.



Evaluating Step 3 (Reliability) means we cannot avoid this selection process, although we may be able to elide them because, depending on the research question and framing, certain steps are theoretically or conceptually unimportant. But as the likelihood of war has declined, Step 3 considerations are less important in assessing alliance efficacy than Steps 1 and 2. In this section, I evaluate two measures of internal alliance health as alternatives partnership reliability.

As depicted in Figure 4, alliances must first solve an internal coordination problem before they can confront external threats. In particular, Cesa (2010) notes that, at minimum, a military partnership requires a single, common animating goal. But allies frequently pursue multiple objectives within and through a single pact. In "homogeneous" partnerships, these additional aims – whether member-specific or

shared across allies – do not interfere with or inhibit pursuit of the pact's central objective. In "heterogeneous" alliances, by contrast, they do. At the limit, these contradictory objectives can undermine internal cohesion and collapse a pact.

We can therefore define internal cohesion as the extent to which states abide by intra-allied agreements and expectations for foreign policy, military, and security coordination. Unlike reliability, cohesion/fragmentation is not conditioned on another event (e.g. failed deterrence and performance opportunity). It is inherent to alliances as institutions. All partnerships require some minimal degree of cohesion to be formed and continue existing. Even as the likelihood of war has declined, cohesion remains a key issue for allies hoping to deter future challenges, enhance collective policy responses through coordination, and develop tighter security communities.

Helpfully, ATOP provides measures of "alliance failure," where members' centrifugal, heterogeneous interests create sufficient fragmentation to collapse the pact.[16] At the alliance-level, *Term* denotes "If the alliance ends due to violation of provisions by one or more members, including willful abrogation before the scheduled termination date..."[17] This measure encompasses performance opportunities, but also includes non-wartime abrogation. The Baghdad Pact, for example, ceased to function as a coherent organization when Iraq and then Iran left due to coups. Neither country abrogated during war, so they are not included by Berkemeier and Fuhrmann (2018) nor Leeds, Long and Mitchell (2000). However, their actions shattered broader American and British ambitions for the pact to serve as a bulwark against the Soviet Union, falling under *Term*.

[16]Indeed, Mattes (2012) uses the first of these measures to assess the conditions under which institutionalization improves alliance cohesion.

[17]Leeds (2018, 20). *Term* as used here refers specifically to Level 2 of the variable.

The first row of Table III presents alliance failure rates based on *Term*. From 1815-2003, only 24.5 percent of alliances, and 9.1 percent of alliance-dyads suffer this kind of failure. The difference between these two values suggests that smaller or bilateral pacts disproportionately suffer from alliance failure. Moreover, these failure rates are generally lower than those for Step 3 alliance violation discussed above. Insofar as allies are concerned not about war, but managing heterogeneous member goals, they can be reasonably confident that their pacts will survive these challenges. Further, as mentioned, cohesion/fragmentation is an "unconditioned," persistent alliance concern. We can construct conceptually coherent, "annualized" version of *Term*, as seen in Row 2. Failure rates are marginal: Year-to-year, alliances are strongly coherent and rarely collapse due to internal policy disagreement. Finally, these rates are even lower following World War 2. This answers the motivating puzzle: We see more and larger alliances post-World War 2 because interstate war has declined and alliances have improved their internal coordination.

Term was measured at the alliance-level, and so it could still mask significant variation in cohesion among allies. As a final check, I use ATOP's *Term Cause* variable, which denotes the project's "judgment regarding why an *alliance member* terminated its affiliation with a given alliance." (Leeds, 2018, 21. Emphasis added.) This allows us to more finely disaggregate the influence of individual defections on alliance failure. I use those values of *Term Cause* encompassing defection due to policy disputes unrelated to alliance management, those related to alliance management, intra-allied military conflict, allied war with a third party that was lost or where allies did not fulfill their obligations, and direct member violations of obligations.[18]

[18]Losing a war with a third party may not reflect internal cohesion, but simply military weakness. However, this only occurs in 61 incidents, and dropping these observations does not substantively

Table III: Alliance Failure Calculations

	All Observations		Post-World War 2	
	Alliance	Alliance-Dyads	Alliance	Alliance-Dyads
Term	0.245 159/648	0.091 400/4389	0.169 78/462	0.058 232/3960
Term (Yearly)	0.018 158/8628	0.062 6470/103658	0.011 78/6819	0.029 2871/69912
Term Cause	0.255 165/648	0.117 512/4389	0.154 71/462	0.082 326/3960
Term Cause (Yearly)	0.019 164/8628	0.053 5458/103658	0.01 70/6819	0.039 3816/69912

This measure includes somewhat more incidents of alliance fragmentation than does *Term*. But, as seen in Rows 3 and 4 of Table III, it provides similar failure rates, both across 1815-2003 and post-World War 2.

Overall, if we believe that states care as much about internal coordination as they do about external deterrence, then the calculations presented above suggest they can be reasonably confident their pacts will survive. At the alliance-level, a failure rate of approximately 25 percent across 1815-2003 tracks with Leeds (2003) fulfillment rate of just under 75 percent. Further, alliance cohesion has improved post-World War 2, contributing to the increasing number and size of military pacts during this era. Failure rates of approximately 16 percent are significantly better than the 78.88 percent calculated by Berkemeier and Fuhrmann (2018). Moreover, failure rates are generally lower than violation rates even when using the same unit of analysis.

change the findings.

Most importantly, the reliability literature's operationalization increasingly diverged from the empirical patterns found in data spanning nearly 200 years. Alliance cohesion/fragmentation, by contrast, accounts for the changing risk of warfare and avoids the data censoring issue, while still providing a theoretically coherent measure of alliance health and strength.

5 Alliance Institutionalization and Reliability

In sum, this study joins Gibler (2006), Poast (2010), and Fordham and Poast (2014) in highlighting how operationalization choices can alter our empirical findings about alliances. To further demonstrate this, this section re-evaluates Leeds and Anac (2005) using these new operationalizations of alliance fulfillment and collapse. The rational design of institutions literature expects formalization and institutionalization to bolster a pact's reliability.[19] Leeds and Anac (2005) find the opposite: These characteristics systematically reduce alliance fulfillment.

The analysis above, however, suggests that using alliances as the unit of analysis may mask the degree to which institutionalization and formalization affect reliability. The former in particular is theoretically associated with *both* increased alliance size and reliability. Coordinating bodies are necessary to manage the policy spillover and issue interdependence that larger membership size generates.[20] These challenges can fragment alliances,[21] such that they require additional costly signals to demon-

[19] Fearon (1997); Koremenos, Lipson and Snidal (2001); Wallander (2000); Bensahel (2007).
[20] Wallander and Keohane (1999).
[21] Cesa (2010).

strate reliability.[22] Bilateral pacts avoid these reliability issues simply due to their smaller size, and so we expect institutionalization to be negatively correlated with bilateralness and the success of bilateral pacts.[23]

In using "alliance" as the unit of observation, we increase the proportion of small partnerships in the dataset, while under-representing multinational pacts relative to their prevalence within the interstate security system, as seen in Figures 1 and 2. Consequently, Leeds and Anac (2005) may simply be picking up bilateral pacts' reduced need for institutionalization as a requirement for success. If institutionalization is indeed associated with alliance violation, then switching to alliance-dyads should buttress their findings: More multinational pacts in the dataset should increase the share of institutionalized alliances. If institutionalization disproportionately causes violation, the coefficient on that variable will remain negative and might be substantively larger than in the original study.

To test this, I create a dataset of alliance-dyads, breaking up all alliances into their constituent interstate links. This produces a 4,389 observation dataset, of which 2,248 experienced a performance opportunity, following Leeds and Anac's coding structure. The dependent variable is *Honor*, following Leeds (2003), Leeds and Anac (2005), and Berkemeier and Fuhrmann (2018). I replicate Leeds and Anac (2005)'s independent variables, of which the two primary ones are:

- *Military Institutionalization*: A trichotomous variable (low, medium, high) based on whether the alliance calls for an integrated command structure, common defense policy, military bases, a formal organization, and other provisions;

[22] Fearon (1997); Koremenos, Lipson and Snidal (2001); Koremenos (2016); Morrow (2000).
[23] Lipson (1991); Wallander and Keohane (1999).

- *Formalization*: A trichotomous variable (low, medium, high) based on whether the alliance calls for ratification, a public declaration, or includes a named coordinating organization;

These are alliance-level variables. I convert them into alliance-dyad ones by following the authors' original coding scheme. If one ally violates its commitment, for example, all alliance-dyads are violated. Similarly, if an alliance is coded as institutionalized, then its alliance-dyads are as well. This should bias the model results in Leeds and Anac (2005)'s favor, as it inflates the frequency of alliance violation and institutionalization in the dataset.[24] However, if institutionalized alliances have more members *and* if they disproportionately honor their commitments, then using alliance-dyads should reveal a positive relationship between military institutions and reliability, one that using alliances as the unit of analysis masks.

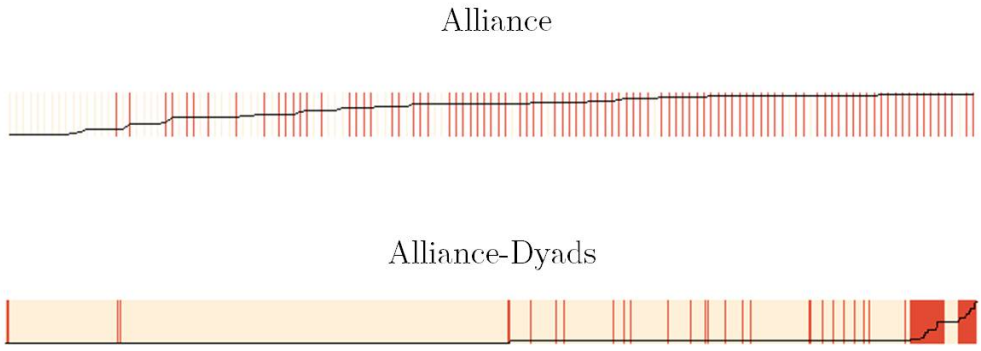
And that is what we see in Table IV. For brevity, the table only presents coefficients and standard errors for *Military Institutionalization* and *Formalization*. Leeds and Anac (2005) is replicated where "Honor" as the dependent variable and "Alliance" is the unit of observation. Both *Institutionalization* and *Formalization* are negative and significant. However, the row immediately below (Honor, Alliance-Dyad) converts Leeds and Anac's data into alliance-dyads. *Military Institutionalization* is now positive and significant, implying that increased military coordination bolster alliance reliability. To obtain substantive quantities of interest, I conduct a 500-run simulation using Zelig. High military institutionalization increases alliance fulfillment by 16.52

[24]To be clear, a strict alliance-dyad coding would not apply the alliance-level variable uniformly to all constituent dyads. Instead, it would measure each member's participation in, say, coordinating organizations, creating more nuanced, dyadic versions of these variables. By contrast, following Leeds and Anac (2005)'s coding scheme generates the maximum number of alliance violations in the conversion process.

percent. Moreover, *Institutionalization* retains its sign, significance, and magnitude when we add censored data back into the analysis (Honor, Alliance-Dyad with Censored Data) and use an annualized version (Honor, Alliance-Dyad-Year). In total, once we adjust for increasing alliance size and longevity, military institutions improve fulfillment, in line with rational institutionalism.

Diagnostics suggest that alliance-dyads more accurately models the data generating process. Figure 5 below displays separation plots for Leeds and Anac's original model and the alliance-dyad version presented here.[25] These plots visualize the empirical distribution of events and non-events compared to their fitted values. Ideally, a well-fitted logistic model should line events on the plot's right-hand side and non-events on the left, perfectly predicting and separating observations based on the covariates.

Figure 5: Separation Plots with *Honor* as the Dependent Variable. Each event marked by vertical red line.



The top graph of Figure 5 displays Leeds and Anac (2005)'s plot. Events and non-events are interspersed together, suggesting that their model did not clearly distin-

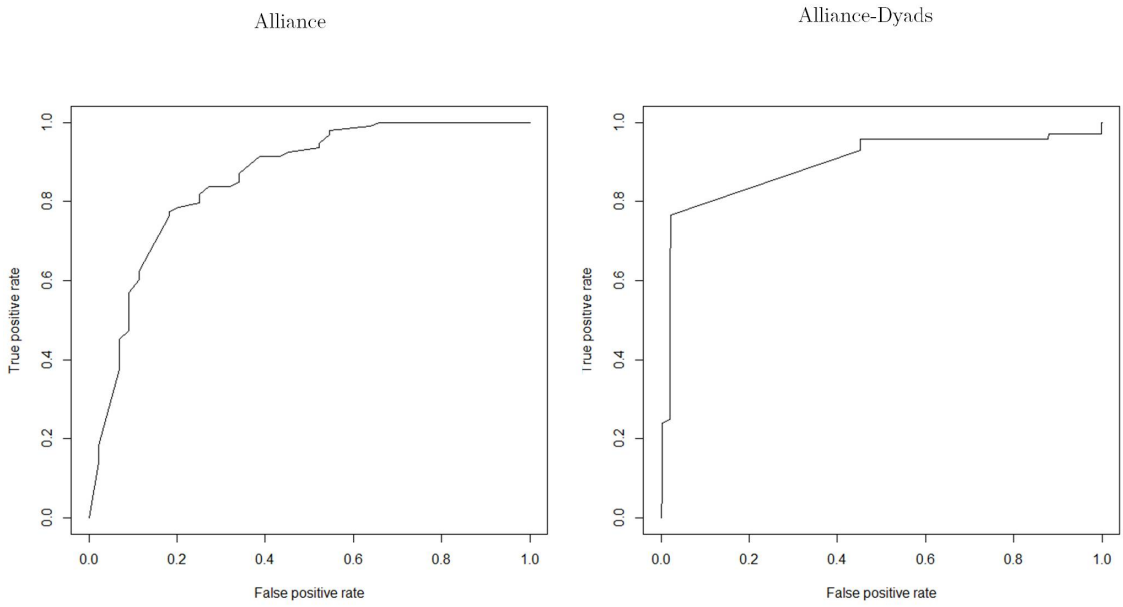
[25]Specifically, "Honor, Alliance" and "Honor, Alliance-Dyad".

Table IV: Results on *Military Institutionalization* and *Formalization* by Unit of Observation, Replicating and Extending Leeds and Anac (2005).

Unit of Observation	Honor				Fail			
	Military Institutionalization		Formalization		Military Institutionalization		Formalization	
Alliance, Data Censored	-0.709	*	-0.929	*	0.159		0.949	*
	(0.298)		(0.359)		(0.251)		(0.303)	
Alliance-Dyad, Data Censored	1.478	*	0.394		1.008	*	0.192	
	(0.219)		(0.263)		(0.218)		(0.24)	
Alliance-Dyad, Data Not Censored	1.801	*	-1.286	*	0.213		-0.899	*
	(0.136)		(0.167)		(0.109)		(0.164)	
Alliance-Dyad-Year, Data Not Censored	1.145	*	-2.002	*	1.412	*	-1.461	*
	(0.119)		(0.183)		(0.028)		(0.039)	

guish between the levels of *Honor*. By contrast, the alliance-dyad approach (bottom graph in Figure 5) shows significantly better separation and model fit, even without adding censored observations back in. This impression is buttressed by Figure 6, displaying ROC curves for the two models. Although similar, the alliance-dyad model has a larger area under the curve, again suggesting that it better classifies outcomes.

Figure 6: ROC Curves with *Honor* as the Dependent Variable.

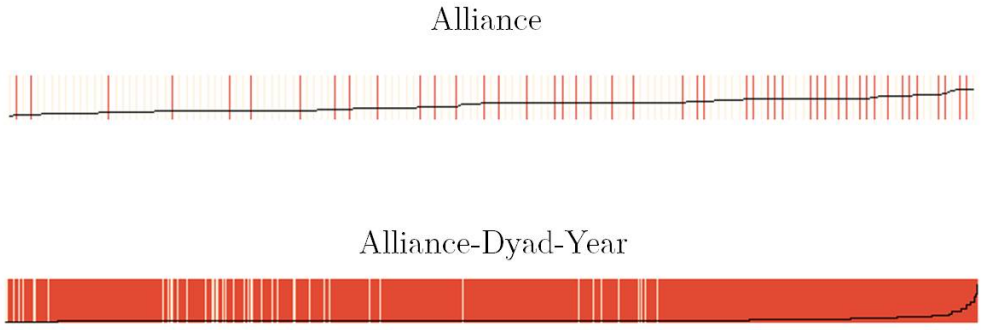


The models do support Leeds and Anac (2005)'s conclusions about *Formalization*. In two of three iterations, that variable remains negative and significant, suggesting that formal pacts increase the likelihood of alliance violation. However, this is substantively less meaningful since performance opportunities have become rarer. Indeed, unreported analysis finds that *Formalization* has no systematic effect on if an alliance experiences an opportunity, while *Military Institutionalization* decreases that chance.[26]

[26]Moreover, again in unreported analysis, the coefficient on *Formalization* actually becomes pos-

Furthermore, *Formalization* also reduces alliance failure, as seen in Table IV's right-hand models.[27] As discussed in the previous section, a conceptually fidelitous measure of alliance cohesion/fragmentation should include observations censored by Leeds and Anac (2005) and be "annualized" to reflect the persistence of this concern within partnerships. This corresponds to "Fail, Alliance-Dyad with Censored Data" and "Fail, Alliance-Dyad-Year" in Table IV. In both, *Formal* is negatively and significantly associated with *Failure*. Simulations suggest that going from a low formality to a high formality alliance decreases the risk of alliance collapse by 30.15%. Separation plots presented in Figure 7 compare "Fail, Alliance" (using the same unit of observation as Leeds and Anac (2005)) with "Fail, Alliance-Dyad-Year." The latter again showing significantly better discrimination between events and non-events. This is also supported by ROC curves (not printed).

Figure 7: Separation Plots with *Fail* as the Dependent Variable. Each event marked by vertical red line.



itive and significant once we correct for spatial interdependence, as we should since alliance-dyads are by definition connected to one another through their security pacts.

[27] *Fail* equals levels 3-7 of *Term Cause*, encompassing alliance collapse due to policy disputes unrelated to the alliance; those related to alliance management; intra-allied military conflict; and/or violation of alliance provisions. Notably, this excludes pacts that end successfully or that are allowed to end.

The models presented above are not meant to be dispositive. The discussion in Section 4 and depicted in Figure 4 suggests that we may need to account for cumulative selection processes to understand the "true" effects of military institutionalization and formalization on alliance fulfillment. Different operationalizations of these concepts, in addition to the unit of observation changes made above, can potentially alter our empirical results further.

However, we can draw three broad conclusions from this section's analysis. First, simply changing the unit of analysis produces substantively different results than those presented in Leeds and Anac (2005). The original analysis is not robust to the long-term changes in alliance membership size, let alone data censoring. Leeds and Anac's negative relationship between *Honor* and *Institutionalization* is likely due to endogeneity between bilateral-ness, institutionalization, and honoring.

Second, diagnostics suggest that the models improve the less they use "alliance" as the unit of observation, at least when using *Honor* or *Fail* as a dependent variable. Disaggregating security pacts into dyads, reintroducing censored data, and even annualization (when conceptually appropriate) all produce better separation between events and non-events. This follows Fordham and Poast (2014) and Poast (2010), which contend that treating multilateral phenomena as single events or bodies misspecifies models and biases estimation.

Third, institutionalization and formalization may potentially have opposing effects. The former bolsters alliance fulfillment, although it may simultaneously spur partnership fragmentation. The latter seems to have the reverse effects, increasing the risk of violations during performance opportunities but improving cohesion as

well.

6 Conclusion

Operationalization matters. Alliances have changed dramatically in size and longevity over the past two centuries. Our selected unit of observation can be more or less aligned with these changes. The latter case can create mistaken inferences driven not by omitted variables or modeling techniques, but simply by conceptualization and operationalization.

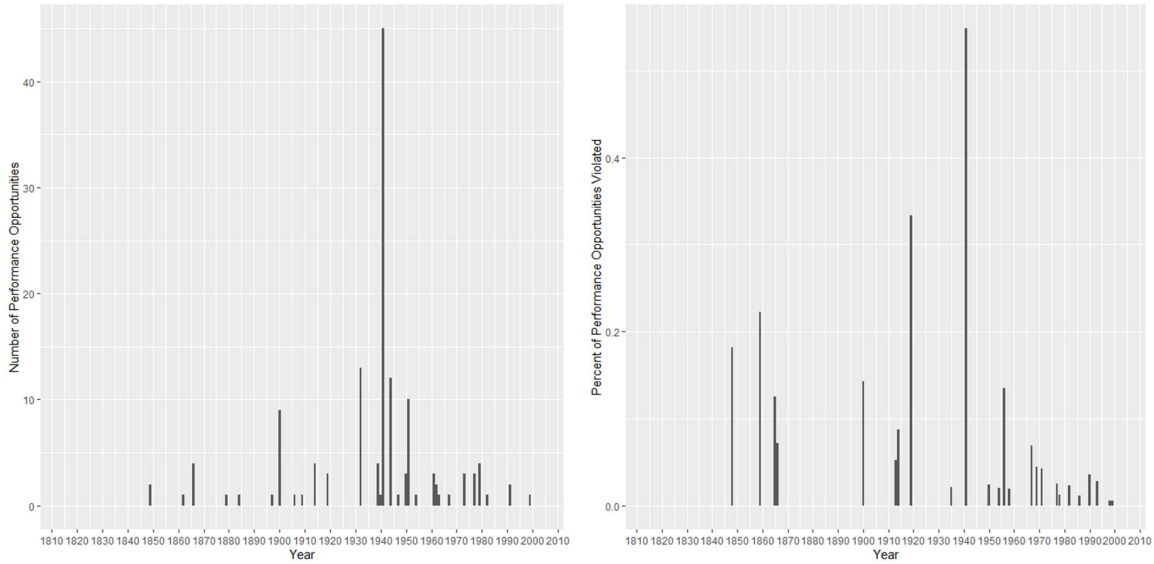
This study makes several points specifically for the alliance reliability literature or quantitative analysis of alliances (whether as a dependent or independent variable) more broadly. First, reliability studies face a tradeoff between external validity and conceptual fidelity. Conditioning on performance opportunities makes conceptual sense, but that operational choice makes it difficult to apply those findings to security partnerships generally.

Second and extended from this, some studies hope to leverage ATOP's entire time frame. In that case, Section 5 suggests that relational and annualized units of observation ("alliance-dyads", "alliance-dyad-year") better reflect the data generating process and lead to superior analytical results than "alliance." There have been far more alliances created in the 59 years from 1945-2003 than in the 130 years prior, and they encompass vastly more interstate security relationships. Dyads or even k-ads better leverage all the information in the ATOP dataset, and findings based on these units of observation suffer from fewer external validity concerns than the "alliance"

unit.

Third, this study has already discussed how conditioning on performance opportunities censors a large portion of the data and introduces external validity concerns. As a final, additional point, opportunities are not evenly distributed, as seen in Figure 8's left-hand graph. They have a "punctuated equilibrium" quality, with stretches of low or no abrogation, followed by isolated spikes of significant defection. These spikes cluster around major wars – especially World War 2 – creating spatial and temporal interdependence. Violations are particularly acute during those conflicts, and much rarer otherwise (right-hand graph in Figure 8).

Figure 8: *Performance Opportunities by Year.*



Insofar as we accept performance opportunities as an analytical condition, these graphs suggest two further areas of research. First, to what extent are performance opportunities during major wars equivalent to those outside of them? Figure 8 suggests they are much more likely to end in violation. Second, why is that so? Is

it simply the magnitude of threat generated by major wars? Or do non-major war performance opportunities erode the alliances that eventually fail during those conflicts? Do we see cascades of abrogation, with alliances "catching" defection from other pacts their members are part of? Returning to Leeds and Anac (2005), to what extent does institutionalization prevent abrogation under these different cases? Regardless, this systematic difference between performance opportunities within and outside of major wars presents an additional problem regarding the extent to which the alliance reliability literature's estimates can be generalized.

References

Bensahel, Nora. 2007. International Alliances and Military Effectiveness: Fighting Alongside Allies and Partners. In *Creating Military Power: The Sources of Military Effectiveness*, ed. Risa A. Brooks and Elizabeth A. Stanley. Stanford, CA: Stanford University Press.

Benson, Brett. 2012. *Constructing International Security: Alliances, Deterrence, and Moral Hazard*. Cambridge University Press.

Berkemeier, Molly and Matthew Fuhrmann. 2018. "Reassessing the Fulfillment of Alliance Commitments in War." *Research and Politics* pp. 1–5.

Cesa, Marco. 2010. *Allies Yet Rivals*. Stanford University Press.

Chiba, Daina, Jesse Johnson and B. A. Leeds. 2015. "Careful Commitments: Democratic States and Alliance Design." *Journal of Politics* 77(4).

Fearon, James D. 1997. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *Journal of Conflict Resolution* 41(1):68–90.

Fordham, Benjamin and Paul Poast. 2014. "All Alliances are Multilateral: Rethinking Alliance Formation in an International System." *Journal of Conflict Resolution* 58.

Gartzke, Erik. 2007. "The Capitalist Peace." *American Journal of Political Science* 51(1):166–91.

Gibler, Douglas M. 2006. The Costs of Reneging: Reputation and Alliance Formation. Presented at the 2006 Shambaugh Conference "Building Synergies: Institutions and Cooperation in World Politics," University of Iowa.

Johnson, Jesse and B. A. Leeds. 2011. "Defense Pacts: A Prescription for Peace?" *Foreign Policy Analysis* 7(1):45–65.

Kim, Tongfi. 2011. "Why Alliances Entangle But Seldom Entrap States." *Security Studies* 20(3):350–377.

Koremenos, B. 2016. *The Continent of International Law: Explaining Agreement Design*. Cambridge University Press.

Koremenos, Barbara, Charles Lipson and Duncan Snidal. 2001. "The Rational Design of International Institutions." *International Organization* 55(4):761–799.

Lascurettes, Kyle. 2020. *Orders of Exclusion: Great Powers and the Strategic Sources of Foundational Rules in International Relations*. Oxford University Press.

Leeds, B. A. 2018. "Alliance Treaty Obligations and Provisions (ATOP) Codebook.". **URL:** *http://www.atopdata.org*

Leeds, B. A. and Burcu Savun. 2007. "Terminating Alliances: Why Do States Abrogate Agreements?" *Journal of Politics* 69(4).

Leeds, B. A., M. Mattes and Jeremy Vogel. 2009. "Interests, Institutions, and the Reliability of International Commitments." *American Journal of Political Science* 53(2).

Leeds, Brett Ashley. 2003. "Do Alliances Deter Aggression? The Influence of Military Alliances on the Initiation of Militarized Interstate Disputes." *American Journal of Political Science* 47(3):427–439.

Leeds, Brett Ashley, Andrew Long and Sara Mitchell. 2000. "Reevaluating Alliance Reliability: Specific Threats, Specific Promises." *The Journal of Conflict Resolution* 44(5):686–699.

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

Leeds, Brett Ashley and Sezi Anac. 2005. "Alliance Institutionalization and Alliance Performance." *International Interactions* 31(3):183–202.

Lipson, Charles. 1991. "Why Are Some International Agreements Informal?" *International Organization* 45:495–538.

Maoz, Z. and B. Russett. 1993. "Normative and Structural Causes of Democratic Peace." *American Political Science Review* 87(3):624–38.

Mattes, Michaela. 2012. "Reputation, Symmetry, and Alliance Design." *International Organization* 66(4):679–707.

Morrow, James D. 2000. "Alliances: Why Write Them Down?" *Annual Review of Political Science* 3:63–83.

Owen, John M. 1994. "How Liberalism Produces Democratic Peace." *International Security* 19(2):87–125.

Poast, Paul. 2010. "(Mis)Using Dyadic Data to Analyze Multilateral Events." *Political Analysis* pp. 403–425.

Sabrosky, Alan Ned. 1980. Interstate alliances: Their reliability and the expansion of war. In *The Correlates of War II: Testing some realpolitik models*, ed. J. David Singer. New York: Free Press pp. 161–98.

Schelling, T. 1966. *Arms and Influence*. Yale University Press.

Schroeder, PW. 1976. Alliances, 1815-1945: Weapons of power and tools of management. In *Historical dimensions of national security problems*, ed. Klaus Eugen Knorr. University Press of Kansas.

Siverson, Randolph M. and Joel King. 1980. "Attributes of national alliance membership and war participation, 1815-1965." *American Journal of Political Science* 24:1–15.

Wallander, Celeste. 2000. "Institutional Assets and Adaptability: NATO after the Cold War." *International Organization* 54(4):705–735.

Wallander, Celeste and Robert Keohane. 1999. Risk, Threat, and Security Institutions. In *Imperfect Unions: Security Institutions Across Time and Space*, ed. Helga Haftendorn, Robert O. Keohane and Celeste A. Wallander. Oxford, UK: Oxford University Press.