I. Introduction

The tau-b and S-score measures have become standard empirical tools for measuring foreign policy similarity (Signorino and Ritter 1999; Häge 2011; Bailey, Strezhnev and Voeten 2017).[1] The tau-b measure on alliance portfolios was first introduced by Bueno de Mesquita (1975) to measure the possibility that two states produce "similar foreign policy responses to war-provoking situations, even though the two nations do not share an alliance with each other," or to "capture the congruence of interests of two states." Since then, the tau-b has been used in a variety of applications, such as determining the tightness of a system (Bueno de Mesquita 1975);[2] calculating the probability of support for adjusted national capabilities after considering alliance effects (Kim 2002); and, perhaps more importantly, as an indicator of the utility of war (Bueno de Mesquita 1981; Bueno de Mesquita 1985; Bueno de Mesquita and Lalman 1992). Signorino and Ritter (1999) suggest, however, that the tau-b does not measure "similarity" but rather the "association" of foreign policies of two states, thus occasionally yielding negative values when it should not. To rectify this issue, they propose the S-score. They also provide insight in recognizing the problem of the no-alliance category in alliance portfolios, and they suggest incorporating UN vote data to the S-score to resolve it.

Despite these advances in alliance measurements, however, few studies have examined their theoretical grounds. This is particularly important given the absence of a "gold standard" for determining whether or not the measurements actually represent the relationships they claim to capture between two states (Bradburn, Cartwright and Fullter

[1] For example, until recently, Signorino and Ritter (1999) was ranked as the fifth most cited paper published in International Studies Quarterly.

[2] He defines systematic tightness as "the degree to which the foreign policies of nations within a single cluster are similar to each other." (Bueno de Mesquita 1975, 199). Tightness is a stronger concept compared to the concept of "interstate interest similarity", because tightness includes the concept of how close the states are in the system as well as whether they share some interstate interests.

2017).[3] In addition, unlike other measurements of political categories, scholars typically use the scores as data in a regression (that is, as a variable), rather than as a statistic. It thus becomes necessary to examine the validity of data generating process for these similarity scores. Furthermore, measurements without theory cannot provide information regarding what exactly is being measured, and interpretation becomes difficult. Without operational interpretations, scholars cannot ensure the comparability of scores across dyads and time, which constitutes the key condition for any valid measure (Anderberg 1973, 77–78). As a result, we are unable to determine the relationship between the measurement and other variables in statistical models, particularly in terms of whether they are confounding, intervening or independent factors (Ray 2005). In short, despite the many insights that the tau-b and S-score measures have generated, *both* suffer from theoretical and interpretational shortcomings. The reliability of the measurements is thus a matter of concern (Hardy and Bryman 2009).

To resolve these issues and to clarify the difference between the existing measures, I introduce an alternative measurement that intentionally approximates both the S-score and tau-b. Using this alternative metric, I reinterpret the tau-like and S-like measures and identify the differences between them. This enhances our substantive understanding of the measures and perhaps offers a path to ensuring the validity of existing studies. In doing so, this paper contributes to a better understanding of alliance measures for use in empirical research designs, thereby improving statistical models in international relations.

The paper proceeds as follows. First, I address the theoretical limitations of the S-score. I then examine Cohen's κ and Scott's π scores. Third, I consider the limitations of tau-b, and I then discuss the construction of a safer measure of alliance similarity that approximates the existing measurements. The fifth section the paper discusses three important, practical uses of the new methods: including foreign policy similarity variables in statistical modeling of alliance formations; reconsidering adjusted national capabilities in light of the alliance effects and; incorporating alliance similarity in k-adic research

[3] Statistically speaking, even hundreds of exemplary cases cannot validate or invalidate the measurements because the number of similarity scores for all available dyads surpasses 600,000.

designs. As an example, Leeds (2003) is replicated., and the final section concludes.

II. Limitations of the S-score

To formulate the S-score,[4] Signorino and Ritter (1999) applied a series of mathematical concepts. In particular, they assumed that the vector of a state's multiple policy choices represents a point in a compact, *N*-dimensional policy space. They also assumed a unit distance (or interval) to all neighboring alliance types or UN vote options. Signorino and Ritter (1999) thus take the data space as a vector space (a ratio scale), and draw from it a metric of normalized absolute distance. The vector space approach implies that the vectors of all states' policy choices share the same origin, such that the values are comparable across all dyads and time.[5] The maximum distances associated with an alliance portfolio and a UN vote portfolio in the S-score setting are set to 3 (between a defense pact and no alliance) and 2 (between yes and no), respectively, representing three times and two times the unit distance of each portfolio. Therefore, before examining their empirical usefulness, it is important to check the validity of the applications of those mathematical concepts in the alliance and/or UN vote portfolios.[6]

First, it may not be possible to impose a ratio scale to alliance portfolios and UN votes. [7] Small and Singer (1969, f.n. 10) noted that alliance types are nominal but also suggested the possibility of ordinality, arguing that the numbers may represent the degree of political

[4] Häge (2011) presented a simplified formula for S: $1 - 2 \times \frac{\Sigma|X_i - Y_i|}{\Sigma d_{max}}$.
[5] A ratio scale must have the same origin.
[6] The S-score also suffers from a computational problem. If the number of countries in calculating the S-score is large (say more than 30), then the S-score converges quickly to +1. For example, the global S-score for the US-UK during the cold war is about +1, but so is the global S-score for the US-USSR dyad. I am grateful to XXXX for raising this issue.
[7] This paper follows Stevens' (1946) levels of measurements. Although Stevens' argument is mainstream in quantitative research, criticisms range from the interpretation of the levels of measurements to his classification itself. However, this paper's theoretical arguments hold regardless of the interpretational differences, and alternative approaches to Stevens' usually add more classifications. Thus, the scale issue is still valid. Regarding interpretations of the levels of measurements, see Hand (1996). For alternative approaches, see (Velleman and Wilkinson 1993); Häge (2011) also distrusts the interval assumption of the S-score.

commitment. Bueno de Mesquita (1975) proposed the loss of autonomy based on what is written in alliance treaties as the reference variable for the rank order of alliance types. Meanwhile, numerous scholars have raised concerns regarding the rank order of alliance portfolios (Wallace 1973; Sabrosky 1980; Levy 1981; Fearon 1997). The general consensus, then, is that alliance portfolios are nominal, or at best, ordinal.

Yet, rather than addressing the criticisms of applying rank order to alliance types, Signorino and Ritter (1999) make the even stronger assumption of a ratio scale in the data. They then introduce other sources of data, in particular UN votes,[8] to solve the problem of the "no alliance" category in alliance portfolios, [9] assuming those data share the same scale. The point is that we cannot simply assume a ratio scale from a nominal or an ordinal scale. If we impose the higher scale on a lower scale, then all variables are taken as ratio scales, but this is not the case (Miller and Salkind 2002, 450). Changing scales thus goes beyond theoretical assumptions and raises concerns about inappropriate conversion.[10] By assuming a ratio scale from ordinal or nominal data, Signorino and Ritter (1999) ignored the theoretical issues involving scale and scale conversion that must be addressed (Hardy and Bryman 2009; Anderberg 1973).[11]

Second, alliance portfolios and UN votes are different in nature as well as scale type. Whereas the UN vote portfolio for a state can be considered a set of policy positions, the alliance portfolio reflects the outcome of those policy positions. And a vote by one player in a potential alliance does not necessarily represent the final outcome. Therefore, even if the scales for alliances and UN votes are converted into ratios, we cannot simply combine them. However, Signorino and Ritter's (1999) approach was to normalize alliance and UN vote portfolios by the maximum difference along each dimension and stack them. Normalizations of this sort are valid only within a portfolio, not across portfolios. The

[8] UN vote data is nominal.
[9] The "no alliance" problem is that there are three instances in the "no alliance" category: no alliance because of hostility, because of irrelevancy to each other's security, and because of an implicit alignment.
[10] According to Anderberg (1973, 53), using a reference variable is "the only available approach to promoting scales." For example, Poole and Rosenthal (1985); Clinton, Jackman and Rivers (2004) make parametric assumptions about the utility function and the error distribution, and establish an interval scale for roll call voting. That is, they take the utility function as the reference variable and promote the scale.
[11] If the S-score is not data but a statistic, we may have relatively generous views on scales (Velleman and Wilkinson 1993).

meaning of 1 in the alliance portfolio is different from the meaning of 1 in the UN vote portfolio, in the same way that measures of 1 kilometer and 1 mile differ.

Third, invoking spatial models, as Signorino and Ritter (1999, p.126) do, exacerbates the problems. To employ a spatial model, we need the concept of utility as a reference variable (Hinich and Munger 1997; Poole and Rosenthal 1985; Poole 2005; Clinton, Jackman and Rivers 2004).[12] Without it, the unit distance assumption within a dataset may not hold. In the case of alliance portfolios, Signorino and Ritter (1999) make the unit distance universal to all dyad-years for all states, such that we can directly compare all states' distances among the alliance types. However, a state's utility from an alliance may change depending on time, strategic situation or regime type. Furthermore, as Lewis and Schultz (2003, 361) point out, "only relative utilities can be inferred" from one player, and it is impossible "to place the utilities of all players on the same scale."[13]

Fourth, the unit distance assumption itself is problematic, independent of the utility argument.[14] Even Tufte (1969, 644–646), who argued that the distinction between interval and ordinal scales is not important if the distinction has no practical bearing on the research, advised against using the approach that Signorino and Ritter (1999) later applied, because it does not "incorporate the researcher's substantive understanding of the thing being measured" and "is not, in any way, a sounder or more conservative choice than any other assignment." If we assign the unit distance to every distance, we sacrifice the advantage of higher scales – the more precise information regarding distances between objects. Moreover, since the unit distance assumption is equivalent to the rank transformation of Spearman, the S-score actually becomes a variant of Spearman's R, contrary to Signorino

[12] The size of the loss of autonomy is also related to utility in terms of a spatial model, because the size determines the preference ordering of a state (Altfeld 1984; Morrow 1991).
[13] Ordeshook (1986, 47-48) neatly explains this issue. Bueno de Mesquita (1975); Altfeld and Bueno de Mesquita (1979) were well aware of the non-comparability of utilities. That is why Bueno de Mesquita (1975, 193) included comparability across nations as a condition for the similarity indicator and Altfeld and Bueno de Mesquita (1979, 96) pointed out that the operationalizations through tau-b "provide a common metric for indicating utility that is monotonic, if not linear, with the utilities actually held by each of the relevant decision makers. Although these interpersonal comparisons are unfortunate, they are necessary if we are to go beyond stating the expected utility model by testing it." Their assumption is minimal to operationalize the indicator in order to test the theory.
[14] For the unit distance assumption, Signorino and Ritter's (1999) rationale was that, "with no other information available, this may be an acceptable scoring rule."

and Ritter (1999).[15]

More importantly, assuming the unit distance without theoretical grounds makes the comparison with tau-b meaningless, because there are "infinitely many underlying frequency distributions for any ordinal distribution" (Hardy and Bryman 2009, 70). Therefore, we may approximate the S-score to tau-b by assigning a proper sized interval between each alliance type. The interval size assumption for approximating tau-b is not inferior to the assumption of unit distance. Rather, this assumption improves on the groundless unit distance assumption by approximating the values to a standard, "tau-b". Furthermore, if scholars desired to assign appropriate intervals using substantive knowledge of each alliance type, it would be practically impossible to do so, as the number of dyad-years in the COW dataset is 657,973.

Fifth, if we cannot compare the distances across different states, then measuring similarity through the concept of distance may pose serious problems. A measure of similarity requires a symmetric relation such that $s_{ij} = s_{ji}$, where s_{ij} represents the similarity between units i and j (Kotz and Johnson 1981, 397). But, if the distances between alliance types differ across states, the requirement of $s_{ij} = s_{ji}$ no longer holds. In that case, we would also not be able to impose the metric on the policy space.[16]

Furthermore, since alliance types are outcomes of joint decisions by allied states, the relationship between states' utility from the alliance types, and the intervals of alliance types, may not be monotonic. If so, even the Pearson's correlation coefficient between utilities and alliance types would no longer be quasi-invariant.[17] In this case, it would be inappropriate to use the S-score to represent utilities of war, because the scores would not reflect the utility of changing alliance types through winning wars.

[15] For a detailed discussion, see the Supplementary Files.
[16] For a metric space to hold, the symmetric condition that $\forall x,y \in X, d(x,y) = d(y,x)$ has to be satisfied.
[17] Since the Pearson's correlation coefficient measures the linear correlation between two variables regardless of their data scale, it is "quasi-invariant to order preserving monotone transformations" (Anderberg 1973, 56).

III. Cohen's κ and Scott's π

Häge (2011) pointed out that the distribution of the S-score's denominator does not reflect the rarity of certain foreign policy ties and states' differential propensities to form such ties. To rectify this, he proposed using Cohen's κ and Scott's π. Unfortunately, those measures also have problems that undermine their suitability as foreign policy similarity measures.[18]

First, negative values on these scales are substantively meaningless; they do not represent the level of disagreement in foreign policy. This issue hinges on the distinction between similarity and agreement. Agreement, a component of similarity, represents decisions that are mutually agreed upon, say, yes or no. Meanwhile, similarity takes into account both agreement and disagreement simultaneously, defined relative to each other. Specifically, if similarity increases, dissimilarity must decrease, and vice versa (Kotz and Johnson 1981, 397-398). The denominator of the S-score, $d_{\max}$, is set to reflect the inverse relationship between similarity and dissimilarity. The S-score is based on a dissimilar measure, $\frac{d_0}{d_{max}}$, proposed by Soergel (401), where d_0 is the observed difference and $d_{\max}$ is the maximum difference.[19] Subtracting it from 1 yields a similarity measure because the denominator is the maximum difference: $1 - \frac{d_0}{d_{max}} = \frac{d_{max} - d_0}{d_{max}}$. The minimum is zero when $d_0 = d_{max}$ (i.e., the total difference), whereas the maximum is 1 if $d_0 = 0$ (no observed difference). On the other hand, κ and π offer no such inverse relationship between agreement and disagreement, because their denominators are the expected differences, not the maximum. Thus, the numerator of the measurements$d_e - d_0$ cannot provide any information about disagreement because the negative values represent the difference between "chance" and "observed", not "agreement" and "disagreement".[20]

[18] If we accept the underlying assumptions of the S-score for the alternatives, all the conceptual problems discussed in the previous section are carried over.

[19] To have the same range as tau-b, 2 is multiplied in the S-score.

[20] κ and π are represented by $1 - \frac{d_0}{d_e} = \frac{d_e - d_0}{d_e}$.

Crystallizing this concern, Krippendorff (1970) wrote that "this makes the coefficient…negative when agreement is *below chance*" (italics mine).[21] Cohen himself also wrote that the negative values of κ do not have any substantive meaning (Cohen 1960, 40).

Second, to use the concept of *chance* agreements, we need the assumptions of independence in units and independence in the decisions of coders (Scott 1955; Cohen 1960). In alliance portfolios, the units are alliances and the decisions of coders represent the leaders' decisions to ally. Those assumptions, however, rule out endogenous relationships and the effect of other alliances in alliance formation.[22] For example, scholars cannot ignore the effect of the Cold War and NATO when evaluating the alliance portfolios of France and Germany during the Cold War period; their decisions to ally correlate closely with those variables. Thus, it is not appropriate to determine a *chance* agreement by multiplying the joint probabilities of each state's proportion of alliances/non-alliances as κ and π have. And if chance agreements are precluded, interpreting the two measures becomes difficult (Uebersax 1987, 140).[23]

Finally, as Häge notes, alliances are rare; κ and π are not reliable for rare events because they are "affected by the prevalence of the *rare* findings under consideration (italic mine)," such that low values may not necessarily mean low rates of overall agreement (Viera, Garrett et al. 2005).[24]

Table 1 displays a hypothetical example: dichotomous alliance ties between states A

---

[21] It may be helpful to consider the original form of the agreement-only measures: $\frac{p_0-p_e}{1-p_e}$, where p_0 represents the observed proportion of agreement and p_e is the expected proportion of agreement. Although Häge argued his measures address the distribution of dissimilarity, this is synonymous with addressing the distribution of agreement because, as Krippendorff (1970, 142–143) showed, Chance-corrected agreement= $1-\frac{d_0}{d_e}$ is equivalent to $\frac{p_0-p_e}{1-p_e}$.

[22] The concept of agreement by chance itself has been criticized, in that the meaning of "by chance" does not represent the random sampling as it appears. See (Pontius Jr and Millones 2011, 4423).

[23] Scott's π has an additional assumption that "the distribution of the entire set of alliances represents the most probable distribution for any individual state" (Scott 1955, 324). This assumption is also not applicable to alliances because it implies that each state's alliance formation patterns are the same as those of all states combined.

[24] Though Viera, Garrett et al. (2005) only mentioned κ, π suffers the same pathology. See fn.21

and B. “Yes” indicates the presence of any alliance commitment between them, and “No” otherwise. The agreement on the alliance ties is high (88%). However, the value of κ is 0.0798 and π is -0.0762, because No prevails. Generally speaking, we cannot say that low values of κ and π represent low levels of agreement in rare instances.

		B		
		Yes	No	Total
A	Yes	**1**	5	6
	No	7	**87**	94
	Total	8	92	100

Table 1: An Example

Introducing κ and π thus seems to create more complex problems rather than solving the issue that Häge raised.

IV. Tau-b: Safer but still problematic

Kendall's tau-b is designed to capture the association between two ordinal (not necessarily interval) variables (Sheskin 2004).[25] It is used to measure the concept that higher ranked objects are preferred to lower ranked objects (Kendall and Gibbons 1990). That is, if the order of an actor’s ranking is known, neither the *scale* of the actor’s utility nor the *distance* between the objects matters.[26] However, while the ranking logic of tau-b is sound, it is also not suitable for alliance portfolios.

[25] See Kendall and Gibbons (1990, Ch.1) for the calculation of tau-b.
[26] However, the fact that the rank number is not important does make it difficult to interpret tau-b in alliance portfolios. This point will be discussed in detail later.

First, general interpretations of tau-b are inappropriate as applications to alliance portfolios. We interpret tau-b as the difference between the probability of concordance pairs and the probability of discordance pairs (Sheskin 2004).[27] For alliance portfolios, the interpretation might be "as the positive value of tau-b increases, the probability that the two states' alliance rank orders are similar increases, whereas negative values of tau-b indicate that the probability of the two states' alliance rank orders being similar decreases." Is this interpretation possible?

The original motivation of rank correlation coefficients was to measure intelligence, which is "a more abstract and higher level concept than specific abilities," by comparing grades of different subjects (Lovie and Lovie 2010). However, a fundamental difference exists between comparing rank-orders in grades and rank-orders in alliance portfolios. Basically, the rank-order of grades always starts from 1st place (with the possibility that it is jointly occupied). In an alliance portfolio, the rank does not necessarily start from 1st place or proceed consecutively. It may start from the second, third, or fourth place, and the ranks are numbered using any number from 1 to 4, ignoring ties. This results in a difference in interpretation of tau-b. Consider the following example. Suppose that states *i*, *j*, and *l* form alliances with the six states A,B,C,D,E, and F as follows. State *i*'s alliance portfolio for states A,B,C,D,E, and F, respectively, is (1,2,2,3,3,3), *j*'s is (0,0,0,0,1,1), and *l*'s is (3,2,3,2,3,2). In that case, the tau-b of states *i* and *j* is 0.64, and the tau-b of states *i* and *l* is -0.40.[28] That is, although states *i* and *j* share alliances in common with only two states (D and E), and although *j* has the lowest level of alliances for the allied states (D,E), if the rank-orders are concordant, tau-b yields a much higher value than that of states *i* and *l*, which share alliances in common with all six states and which have high degrees of commitment with all of them (defense pacts or neutrality). In other words, unlike the case of grades, high tau-b values for alliance portfolios cannot be interpreted as high similarity between the two states' alliance policies.

Second, the calculation of expected utility of war, using "the similarity in policy

[27] In tau-b, concordance means that pairs of ranks from each subject have the same sign under comparison; otherwise, they are discordant.
[28] The tau-b of *j* and *l* is 0.

responses to war provoking situations" may be more complex than Bueno de Mesquita (1981, 109-118) originally considered.[29] Although state *i* wages war against state *j* because *j* has dissimilar policy responses to war, it may not be easy for *i* to change *j* 's alliance types to make them similar to its own alliance preferences among the states. In general, alliance types are determined not unilaterally but mutually by allied partners. Therefore, even if *i* wins the war against state *j*, unless the states allied with *j* share similar alliance type preferences with *i*, or *i* defeats every ally of *j*, *i* may not be able to change *j* 's alliance types according to its own alliance type preferences. State *i* may force *j* to terminate an alliance with a state whose interests differ from those of *i*, or to form an alliance with a state that *i* supports. However, *i* may not be able to designate what kind of alliance type the states form in general, because that would be a joint (or strategic) choice between state *j* and its potential alliance partners.

Third, and following from the first two points, we know that we cannot take a single value (sign and magnitude) for tau-b. In fact, a single value for tau-b does not have any meaning to the authors who introduced tau-b to war study (Bueno de Mesquita 1975; Altfeld and Bueno de Mesquita 1979; Bueno de Mesquita 1981). Instead, the *relative difference* between tau-b values such as $U_A^i - U_B^i$ is essential to the application of tau-b.[30] However, many scholars have considered a single value of tau-b as meaningful, which causes problems in the interpretation of tau-b. A prominent example is to take tau-b as the probability that two states in question act as if they have a defense pact between them.[31] A concrete illustration may provide a clearer picture. Suppose *i* and *j* are allied with states A, B, and C. The alliance portfolios for *i* and *j* with A, B, and C are (1,2,3) and (3,2,1), respectively.

[29] This issue is not limited to tau-b. Any foreign policy similarity measure used to represent utility of war has the problem.

[30] The sign is only meaningful after taking the difference. In addition, if tau-b or the S-score is used as an indicator of utility, only orders of the scores within a state may matter, as utilities are invariant across affine transformations.

[31] This interpretation was used in the calculation of the national capabilities taking into account the alliances, which considered the tau-b values as a sort of "probability of third party support in war" (Organski and Kugler 1981; Kim 1991). This argument also holds for the S-score because what the S-score measures is policy similarity, not the probability that both states' policies are a defense pact.

In the example, if B is threatened, *i* and *j* will adopt the same response (remain neutral). B also responds the same to whomever is threatened between *i* and *j*. But, tau-b remains -1 because the discordant orders dominate. The meaning of discordant orders is that if any of *k*, *i* and *j* are threatened, their reactions will differ. For example, *i* and A have a defense pact, so they defend each other when either of the two of them is threatened.

States	(A,B)	(A,C)	(B,C)
i	(1,2)	(1,3)	(2,3)
j	(3,2)	(3,1)	(2,1)
Score	-1	-1	-1

On the other hand, *j* and A maintain an entente. So, *j* or A's reaction will follow entente if any of *j* and A is threatened. That is, what a single value of tau-b between *i* and *j* represents in the example is whether or not *i* and *j* adopt a similar response to a common allied partner under threat. In that sense, a single value of tau-b ignores two points in the example: the same reactions among states exist even though tau-b is -1 (for B), and all the states are allied with *i* and *j*. Therefore, a single value of tau-b cannot be the probability that both states' policies represent defense pacts, nor can it properly represent the shared interests among states.[32]

V. Measuring Interstate Interest Similarity

[32] Therefore, the adjusted national capabilities proposed by Kim (2002) are dubious. He took the similarity measure difference between the two states in question, $\frac{U_{ki} - U_{kj}}{2}$, as the probability of support based on the problematic interpretation of "acting as if the states have defense pact". If we still want to use $\frac{U_{ki} - U_{kj}}{2}$, we need a function that maps the difference onto support probability. Furthermore, he also used a single value of the similarity measures, $U_{ki} > 0$, to reflect the possibility of a third party's intervention, which is also problematic.

V.1 Alliance Portfolios Are Still Valuable

Some scholars use UN General Assembly (UNGA) vote data to measure foreign policy similarity. For example, to recover the ideal points of states for each issue in the UNGA, Bailey, Strezhnev and Voeten (2017) employ a method similar to the IDEAL method used for the analysis of roll-call votes in the U.S. Congress (Clinton, Jackman and Rivers 2004). However, some limitations persist in using UN vote analysis for high political issues.

First, the estimated ideal points of states depend on the issue at hand. Since UNGA resolutions are nonbinding, those issues may be less important. Indeed, the most important issues, including military security, are dealt with in the UN Security Council (UNSC), where the resolutions are binding, so strategic and/or block voting occur. Therefore, estimates of states' ideal points may not be appropriate for high political issues such as military security, alliances, or very critical non-military issues typically addressed in the UNSC.

Second, a trade-off may exist between the importance of an issue and the sincerity of a state's vote in the UNGA. The nature of UNGA voting differs from roll-call voting in Congress: members of Congress care about their own constituencies and are relatively free from the other members' pressure. However, in the UNGA, states may have to give heed to the powerful states' preference - this is particularly true if the issue up for a vote is salient to the powerful states, such as China's human rights record. Therefore, in the UNGA the validity of the sincere voting assumption may not hold for important issues.

Yet it remains important to measure foreign policy similarity through alliance portfolios in a manner that accounts for high political issues.[33] Doing so requires different methods, depending on the nature of the issue.

V.2 Constructing Interstate Interest Similarity Measures

[33] The time span that UN vote data covers is also limited.

To summarize the discussion to this point, measuring shared interests between states represented by alliance types presents numerous challenges. What can be surmised from alliance portfolios is that, at best, if an alliance exists, the states have some shared interest, regardless of the alliance type.[34] It is thus appropriate and safe to develop a less ambitious indicator using a binary coding scheme.

The coding rule for the similarity coefficient of interstate interests assigns 1 for allied and 0 otherwise. With a total sample size of *N*, we have a 2×2 table with entries *a* (the number of allied states common to both states), *b* (the number of states allied only to state A), *c* (the number of states allied only to state B), and *d* (the number of non-allied states for both states).[35]

		j = 1	*j* = 0	
i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n=a+b+c+d

Table 2: T*he 2*x*2 table of agreements and disagreements*

Since numerous coefficients exist for measuring similarities in interstate interests, some criteria must be established for choosing the optimal measurement for dichotomous variables. Ideally, the measurement should (1) be operationally interpretable, (2) be balanced between matches (agreements) and mismatches (disagreements), and (3) produce reasonable values. The operational interpretability is most important. If a measurement is not interpretable, the values it produces are not comparable (Anderberg 1973, 77–78).

[34] Even if states do not have a shared interest or conflicting interests, the difference may be small enough to be overcome by policy concessions. States can thus come to share some interstate interests.
[35] This coding rule is general for dichotomous variables (Kotz and Johnson 1981, 398).

Focusing only on matches would not be desirable. The coefficients may yield high similarity values for an alliance even if dissimilarities larger than the similarity exist between the two states. Indeed, tau-b and the S-score both take into account matches and mismatches. Per those criteria, two candidates exists for the similarity measure: Pearson's φ and Hamann's measure (or H-score here) (Kotz and Johnson 1981, 398–400). Table 3 provides the formulas for the two measures.

Pearson's ϕ	Hamman (H-score)
$\frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$\frac{(a+d) - (b+c)}{a+b+c+d}$

Table 3: *The Two Candidates for the Similarity Measure*

Meanwhile, for those interested in measuring the tightness between states, I have developed the I-score. The I-score is also based on a dichotomous method. If two states have a common type of alliance, it is coded as 1, and -1 otherwise. Mutually non-allied pairs of states are also coded as 1, and if a state forms an alliance with only one of the two states in question, it is coded as -1.

$$x_{kl} = \begin{cases} +1 & if \quad a_k^i = a_k^j \\ -1 & if \quad a_k^i \neq a_k^j \end{cases} \tag{1}$$

Thus, I-score is given by

$$\frac{\sum_{k,l} x_{kl}}{\sum_{k,l} |x_{kl}|} \qquad k,l = 1,2,\ldots,N; k \neq l \qquad (2)$$

Hamann's method is similar to Wallace's (1973) similarity coefficient φ.[36] The H-score can be interpreted as the frequency of agreements minus the frequency of disagreements. The correlation coefficients between the unweighted H- and S-scores are 0.9801and 0.9202 for regional and global, respectively. The H-score and tau-b's correlation coefficients are 0.6378 and 0.4150 for regional and global, respectively. Like *φ* and tau-b, the H-score can successfully approximate S-score.

Pearson's *φ* is the binary version of the Pearson's correlation coefficient (Anderberg 1973, 85–86). More importantly, however, Pearson's *φ* is exactly the same as the binary version of tau-b (Kendall and Gibbons 1990, 48–49).[37] Since Pearson's *φ* provides a more intuitive interpretation, I adopt it. Pearson's *φ* is a measure of a linear correlation of whether state B also forms alliances with state *k* when state A has alliances with *k*: +1 indicates a perfect linear correlation between state A and state B's alliance formation tendencies. The value -1 means one state's allied states are perfectly mismatched with those of the other state. Their alliance partners do not overlap at all if *φ* is -1. The value 0 indicates no linear correlation between the alliance formation tendencies of the two states.

The correlation coefficients of tau-b and *φ* are 0.9984 for regional, and 0.9997 for global, suggesting that alliance portfolios do not create meaningful differences from a dichotomous categorization of alliances.[38] The means and the standard deviations of *φ* and tau-b are almost identical.[39] Therefore, the binary measurement *φ* can successfully

[36] Though Wallace's φ has the same name as Pearson's φ, φ usually refers to Pearson's φ here; Wallace (1973), however, used the coefficient α that only considers matches.
[37] I cannot find any literature that articulates the equivalency of the two measures in a binary case.
[38] In EUgene, we should consider that, for global unweighted alliance portfolios, observations are missing in about 71% of mutually no alliance cases and in about 75.1% of the total number of states.
[39] See "Supplementary File".

approximate a 4 × 4 rank order measurement of tau-b and replace it with an operational interpretation. Existing studies' results using tau-b may be fundamentally unchanged, though the interpretation and meaning of the similarity measures would differ.

The I-score can be interpreted as "the probability that the two states maintain identical foreign policy." The correlation coefficient of the I-score and tau-b is 0.62 for regional data and 0.39 for global data. However, the correlation coefficient of the I-score and the S-score is 0.98 and 0.92 for unweighted regional and global, respectively. For capability-weighted alliance portfolios (Signorino and Ritter 1999, 124), the correlation coefficients of the two are 0.96 for both regional and global. This indicates that the I-score constitutes an effective dichotomous approximation of the S-score.[40] This approximation works because mutually non-allied states are such a common outcome and the majority of alliance types are defense pacts.

Table 4 provides a concrete example, in which entry letters and ± 1 can be plugged into the table for the formulas for H, ϕ and I. In the example, out of 10 states, *i* and *j* share 5 mutually allied states and 1 mutually non-allied state, whereas 3 states are allied with only one of them. Thus, we can see that *i* and *j* share at least a moderate level of interest. The S value, 0.2, then, seems too low to capture those interests. Note that due to the unit distance assumption, the distances between no-alliance and entente and other neighboring alliances are the same, which ignores the rarity of alliances in reality. The H-score, which captures the marginal frequency of agreements (states sharing interests in common with *i* and *j*), is 0.4, a moderate level of interest. Meanwhile, the I-score is -0.6 because it focuses on "the probability that the two states produce the same policy in war-prone situations" rather than "shared interests". The reason tau-like measures are different is that ϕ captures the linear correlation of the "form-or-not" pattern, whereas tau-b measures "alliance type" patterns.[41] In general, if the mutually allied types have similar portfolios, including non-alliances, and the total number of states in the dataset increases, the S-score and S-like measures converge

[40] The correlation coefficients of the I-score and the H-score are also high, resulting in good approximations for each other. The correlation coefficients of weighted H and I are 0.9871 and 0.9878 for regional and global, respectively.
[41] For the tau-b calculation, see "Supplementary File".

to similar values. For ϕ, as the alliance types become more concordant, ϕ converges to tau-b.

	i	j	s	Entry	$X_i = Y_i$
i	3	0	0.2	b	-1
j	0	3	0.2	c	-1
A	0	0	0	d	1
B	3	2	0.067	a	-1
C	1	2	0.067	a	-1
D	2	3	0.067	a	-1
E	1	2	0.067	a	-1
F	1	2	0.067	a	-1
G	1	0	0.067	b	-1
H	3	3	0	a	1

S-score	0.2	$s = \frac{2\lvert x_i - y_i \rvert}{30}$
H-score	0.4	$S = 1 - \Sigma s$
I-score	-0.6	Entry: as in Table 2.
ϕ-score	0.22	$X_i = Y_i \rightarrow \ 1$
Tau-b	0.12	$X_i \neq Y_i \rightarrow -1$

Table 4: A Concrete Example

V.4 Three important, practical issues related to foreign policy measurements.

Now consider three important, practical issues related to alliance similarity in existing empirical studies.

V.4.1 Why alliance similarity measures should be included in statistical analysis for alliance formation

Lai and Reiter (2000) and Fordham and Poast (2016) argued that alliance similarity is a tautological variable in the prediction of alliance formation. This, however, may not necessarily be the case. Here, I consider this criticism in light of the methods proposed in this paper. The operational interpretation of the methods proposed in this paper is basically the marginal frequency of matches (or interests) shared by the two states or a linear correlation of two states' alliance formation patterns. In fact, states do not necessarily make an alliance even if such shared interests are great, because straightforward alignment is possible. There is also no guarantee that the two would not form an alliance if their shared interests measured by alliance similarity are small, because policy concessions may contribute to forming an alliance (Morrow 1991). Moreover, according to the size principle (Riker 1962), an increase in the number of shared alliances of the two states does not necessarily lead to the formation of an alliance between them. Furthermore, it is possible to control for a cross-sectional correlation in alliance similarity by controlling for the marginal frequency of matches in the alliance similarity pattern. Therefore, the alliance formation similarity is a variable that must be controlled for or tested in the study of alliance formation as opposed to Lai and Reiter (2000) and Fordham and Poast (2016).

V.4.2 Reconsidering adjusted national capabilities in light of alliance effects

Since it is difficult to interpret the tau-b and S-score measures as the utility of war in a system, it is necessary to reconsider those interpretations in adjusted national capabilities. (Kim 2002) has proposed a way of calculating adjusted national capabilities that combines internal and external capabilities.[42]

[42] The concept of adjusted capabilities addressed here has been improved from (Kim 1991). Originally, there was no such condition as $U_{ki} > 0$. The new condition intends to reflect the possibility of a third party's noninvolvement.

$$NC_i = IC_i + EC_i$$

where

$$EC_i = \sum_{k \neq i,j} IC_k \times (U_{ki} - U_{kj})/2 \qquad \text{if} \qquad U_{ki} \geq U_{kj} \quad \text{and} \, U_{ki} > 0,$$

where i is nation i in each dyad and k is a third party; NC_i is i's adjusted national capabilities; IC_i is i's internal capabilities; EC_i is i's external capabilities; IC_k is k's internal capabilities; U_{ki} (or U_{kj}) is the shared interest of k attached to k.[43]

The term $(U_{ki} - U_{kj})/2$ represents the probability that k will support i in case of conflict between i and j. However, there are significant limitations in the interpretation of the term as a measure of the probability of support.

First, if we interpret the term $(U_{ki} - U_{kj})/2$ based on "alliance type similarity", then, as noted, U_{ki} (or U_{kj}) would mean the probability that the two states k and i (or j)'s reactions to war-prone situations are identical. The identical policy response does not necessarily mean that they react as if there is a defense pact between them. Therefore, $(U_{ki} - U_{kj})/2$ contains all the probabilities of three types of alliance behaviors such that it cannot be taken as the marginal probability of intervention of k for i as in the equation.

Second, if we use the H-score, the term must be interpreted based on "alliance formation pattern similarity". Then we can define utility as the marginal similarity in an alliance formation pattern. However, in this case, the term $(U_{ki} - U_{kj})/2$, cannot be taken as the difference between the k's probabilities of support for i and j. Instead, we need a function that maps "the difference in utilities" onto "the difference in support probability." The relationship between utility and support probability might be a weakly monotonic relationship. However, there is no reason to believe that the function is $f(U_{ki} - U_{kj}) = (U_{ki} - U_{kj})/2$, where f is the function that maps utility onto support probability.

Third, from the equation, we have $(U_{ki} - U_{kj})/2$ if $U_{ki} \geq U_{kj}$ and $U_{ki} > 0$. In the case that $U_{ki} > 0$ but $U_{kj} < 0$, as noted, it is possible that tau-b or S-score can yield negative numbers when there are actually shared interests between the two states k and j. Therefore, it would be

[43] (Kim 2002) used S-score to calculate U_{ki} (or U_{kj})

incorrect to state unequivocally that a negative tau-b or a negative S-score means that *k* would not help the other states. Thus, using similarity scores to represent the probability of support has serious limitations and cannot be trusted as an accurate measure.

V.4.3 Similarity measures in a k-adic research design

Poast (2010,2016) propose a k-adic research design to deal with multilateral events that have been overlooked in dyadic research designs. Fordham and Poast (2016) applied a k-adic research design to investigate multilateral characteristics in alliance formation. To control for the heterogeneity of the prospective alliance partners, they included the average S-score of all the dyads within the k-ad. However, they did not provide a rationale for using the average S-score.

If we want to represent the similarity values across the states in a k-ad, there can be better alternatives based on theoretical and/or empirical grounds than the average S-score. In other words, in order for k-ads to be established as a major research design in international studies, many measurements and indices developed for use with dyads have to be newly developed or redefined for k-ads, and this is equally true for the alliance similarity measure.

This paper proposes three possible candidates for k-adic similarity measurement, along with Fordham and Poast (2016)'s simple average: 1) using the value of similarity measures between a state of interest and the largest state in the k-ad, 2) using the weighted average of the similarity values of all the dyads within the k-ad, and 3) using the weighted median of the similarity values of all the dyads within the k-ad. Using similarity values between a state of interest and the largest state of a k-ad can be compared to a measure of systemic dissatisfaction in Power Transition Theory, because systemic dissatisfaction is measured by alliance similarity between a state of interest and a dominant state (the largest state) (Kim 1991).[44] The weighted average would be better than the simple average that Fordham and Poast (2016) used because it takes the capabilities into account in calculating the average, and if it is properly weighted, it also represents a powerful tool for predicting actual cases as used in the Predictioneer's Game (Bueno de Mesquita 2009, 2011). The weighted median value is also very good candidate, because the weighted median can invoke theoretical grounds such as the median voter theorem (Downs

[44] Poast (2010) also assumed that "the largest state's capabilities relative to the entire group's capabilities" are major factors in a decision to join a multilateral alliance or not.

1957), and a prediction model has also proved its usefulness as a tool for predicting actual situations (Bueno de Mesquita, Newman and Rabushka 1985, 40). Which one would be the best to capture similarities within a k-ad may depend on the research topic and empirical results.

Here it is essential to recognize that these are important areas in which to test alternative indicators. With that in mind, I now turn to a replication analysis to show the feasibility of using the proposed alternative indicators.

V.5 A Replication of Leeds (2003)

A replication of the analysis in Leeds (2003) illustrates how the new measures presented in this paper are different from and/or similar to each other and the existing measures. Leeds (2003) serves as an appropriate replication case for the following reasons. First, the S-score in Leeds (2003) was generated by an earlier version of EUgene. However, after Leeds (2003), at least three major updates to the S-score have taken place that could affect the results of empirical studies.[45] In fact, the S-score used in Leeds (2003) differs significantly from the updated S-score: the correlation coefficient between the two is only 0.6343. Out of 69,836 observations in total, 68,474 observations (98%) differ across the old and new S-scores. 3,730 observations (5%) have different signs. Observations even exist for which the old S-score is greater than 0.5 but the new one is less than zero, and vice versa (308 cases). Therefore, empirical results may clearly differ when the updated S-score is used. Second, for the samples used in Leeds (2003), the unweighted global alliance portfolio does not include missing data, which enables a comparison between global tau-b and global φ. We can thus compare every global measurement at once.

	S-score	I-score	H-score	Tau-b	φ
S-score	1.0000				
I-score	0.9221	1.0000			
H-score	0.9137	0.9779	1.0000		

[45] For updated information, see http://eugenesoftware.org.

Tau-b	0.3738	0.3519	0.4215	1.0000	
φ	0.3715	0.3503	0.4228	0.9996	1.0000

Table 5: *The Correlation between Similarity Measures*

However, I use a different statistical technique than Leeds (2003) did. Leeds (2003) used a generalized estimating equation (GEE), which can control for autocorrelation as well as cross-sectional correlations.[46] However, to correct for autocorrelation, Leeds (2003) used an "exchangeable" correlation structure that assumes that the observations of a dependent variable within a dyad are all equally correlated across time. Like the homogeneity variance assumption in the OLS model, the "exchangeable" assumption is unrealistic, and thus not suitable for dispute initiation. The problem is that if we use GEE, the effect of a defense pact on dispute initiation, the core independent variable, estimated to be significant at the 5% level in Leeds (2003), is not statistically significant at any conventional level for any model except the old S-score model. That is, Leeds (2003)'s results hinge almost entirely on the old S-score.[47] This paper uses the technique proposed by Carter and Signorino (2010) instead of GEE. In particular, t, t_2, and t_3 are included in the logit regression controlling cross-sectional correlation (dyads), where t represents the time since the last dispute between the states in a dyad was observed.

Table 5 shows that, for all similarity measures, all variables of interest have the expected signs and are statistically significant at the 1% level. Joint Democracy, which was not significant in Leeds (2003), also has the expected sign and is statistically significant at the 5 % level for all similarity measures.

Regarding the similarity measures, tau-b and φ are still very similar. The S- and H-scores are more similar than the S- and I-scores. Furthermore, the S-score is not significant at any conventional level, whereas the I-score is significant at the 5% level and the H-score is

[46] See Zorn (2001).
[47] For the full results, see the "Supplementary File".

significant at the 10% level. Those results suggest that even though the overall correlation is higher between the S-score and the I-score than between the S-score and the H-score, in specific statistical analyses the relationship can be reversed. Furthermore, the statistical significance also varies across analogous similarity measures. Therefore, although the alternative measures may successfully approximate the existing ones, it is important to replicate those studies as a validity check.

V. Conclusion

This paper offers an enhanced understanding of foreign policy similarity measures and theoretically safe alternatives. Nevertheless, it is important to note that alliance portfolios themselves have a selection bias problem. In particular, the global index includes too many irrelevant mutually non-allied states. For the regional index, if two states in question belong to different regions, selection bias also occurs. Therefore, opportunities remain for correcting the selection bias issue in both global and regional alliance portfolios.

That said, the paper underscores the idea that there may be no such thing as a universally "generalizable" similarity measure. Depending on the research topic, researchers may need different measures. The point to take away is that the greater the understanding of similarity measures and our own research topics, the more accurate and richer the measurements of similarity and the research results will be.

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Joint Democracy	-0.518**	-0.479**	-0.527**	-0.546**	-0.541**	-0.518**	-0.518**
	(0.241)	(0.240)	(0.239)	(0.241)	(0.240)	(0.241)	(0.241)
Contiguity	0.879***	1.004***	0.939***	0.972***	0.954***	0.880***	0.880***
	(0.128)	(0.131)	(0.128)	(0.127)	(0.126)	(0.127)	(0.127)
Power of Potential Challengers in Relation to Potential Target	0.522***	0.533***	0.528***	0.530***	0.527***	0.522***	0.522***
	(0.132)	(0.132)	(0.132)	(0.131)	(0.132)	(0.132)	(0.132)
Shared Alliance Commitment	-0.255*	-0.137	-0.208	-0.202	-0.198	-0.244	-0.242
	(0.152)	(0.157)	(0.159)	(0.158)	(0.161)	(0.253)	(0.250)
S-score (Old) S-score (New)		-0.535***					
		(0.116)					
			-0.235				
			(0.145)				
I-score				-0.291**			
				(0.125)			
H-score					-0.239*		
					(0.127)		
Tau-b						-0.0157	
						(0.242)	
ϕ							-0.0183
							(0.237)
Potential Target Has Defensive Ally	-0.377***	-0.488***	-0.408***	-0.421***	-0.413***	-0.377***	-0.376***
	(0.112)	(0.115)	(0.115)	(0.115)	(0.116)	(0.112)	(0.112)
Potential Challenger Has Offensive Ally	0.550***	0.471***	0.506***	0.496***	0.507***	0.549***	0.549***
	(0.111)	(0.111)	(0.116)	(0.114)	(0.115)	(0.111)	(0.111)
Potential Challenger Has Relevant Neutrality Pact	0.560***	0.515***	0.566***	0.556***	0.559***	0.560***	0.560***
	(0.0985)	(0.0993)	(0.0985)	(0.0979)	(0.0981)	(0.0985)	(0.0985)
Peace Years	-0.113***	-0.106***	-0.113***	-0.113***	-0.113***	-0.113***	-0.113***
	(0.0101)	(0.0104)	(0.0101)	(0.0101)	(0.0101)	(0.0101)	(0.0101)
Peace Years^2	0.00178***	0.00166***	0.00178***	0.00178***	0.00178***	0.00178***	0.00178***
	(0.000250)	(0.000254)	(0.000252)	(0.000252)	(0.000252)	(0.000250)	(0.000250)
Peace Years^3	-7.66e-06***	-7.10e-06***	-7.72e-06***	-7.71e-06***	-7.71e-06***	-7.66e-06***	-7.66e-06***
	(1.67e-06)	(1.68e-06)	(1.68e-06)	(1.69e-06)	(1.69e-06)	(1.67e-06)	(1.67e-06)
Constant	-3.991***	-3.763***	-3.887***	-3.887***	-3.903***	-3.991***	-3.991***
	(0.157)	(0.157)	(0.174)	(0.168)	(0.170)	(0.157)	(0.157)
Observations	69,730	69,730	69,730	69,730	69,730	69,730	69,730

Robust standard errors in parentheses
**** p<0.01, ** p<0.05, * p<0.1*

Table 6: *The Replication of Leeds (2003) (New Method)*

References

Altfeld, Michael F. 1984. “The decision to ally: A theory and test.” *The Western Political Quarterly* pp. 523–544.

Altfeld, Michael F and Bruce Bueno de Mesquita. 1979. “Choosing sides in wars.” *International Studies Quarterly* pp. 87–112.

Anderberg, Michael R. 1973. *Cluster analysis for applications*. Academic press.

Bailey, Michael A, Anton Strezhnev and Erik Voeten. 2017. “Estimating dynamic state preferences from United Nations voting data.” *Journal of Conflict Resolution* 61(2):430–456.

Bradburn, Norman M., Nancy L. Cartwright and Jonathan Fullter. 2017. A Theory of Measurement. In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, ed. Leah McClimans. 1 ed. London: Rowman Littlefield International chapter 5, pp. 73–88.

Bueno de Mesquita, Bruce. 1975. “Measuring systemic polarity.” *Journal of Conflict Resolution* 19(2):187–216.

Bueno de Mesquita, Bruce. 1981. *The war trap*. Yale University Press.

Bueno de Mesquita, Bruce. 1985. “The war trap revisited: A revised expected utility model.” *The American political science review* pp. 156–177.

Bueno de Mesquita, Bruce. 2009. *The Predictioneer’s Game: Using the logic of brazen self-interest to see and shape the future*. Random House Incorporated.

Bueno de Mesquita, Bruce. 2011. “A new model for predicting policy choices: Preliminary tests.” *Conflict Management and Peace Science* 28(1):65–87.

Bueno de Mesquita, Bruce and David Lalman. 1992. *War and reason: Domestic and international imperatives*. New York:Cambridge Univ Press.

Carter, David B and Curtis S Signorino. 2010. “Back to the future: Modeling time dependence in binary data.” *Political Analysis* 18(3):271–292.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. “The statistical analysis of roll call data.” *American Political Science Review* 98(2):355–370.

Cohen, Jacob. 1960. “A coefficient of agreement for nominal scales.” *Educational and psychological measurement* 20(1):37–46.

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. 2016. “All alliances are multilateral: rethinking alliance formation.” *Journal of Conflict Resolution* 60(5):840–865

Häge, Frank M. 2011. “Choice or circumstance? Adjusting measures of foreign policy similarity for chance agreement.” *Political Analysis* 19(3):287–305.

Hand, David J. 1996. “Statistics and the theory of measurement.” *Journal of the Royal Statistical Society. Series A (Statistics in Society)* pp. 445–492.

Hardy, Melissa A and Alan Bryman. 2009. *Handbook of data analysis*. Sage.

Hinich, Melvin J and Michael C Munger. 1997. *Analytical politics*. Cambridge University Press. Kendall, Maurice George and Jean Dickinson Gibbons. 1990. *Rank correlation methods*. Edward Arnold. Oxford University Press.

Kim, Woosang. 1991. "Alliance transitions and great power war." American Journal of Political Science pp. 833–850.

_____. 2002. "Power parity, alliance, dissatisfaction, and wars in East Asia, 1860-1993."
Journal of Conflict Resolution 46(5):654–671.

Kotz, Samuel and Norman L Johnson. 1981. *Encyclopedia of statistical sciences,vol.5.* Wiley Online Library.

Krippendorff, Klaus. 1970. "Bivariate agreement coefficients for reliability of data." *Sociological methodology* 2:139–150.

Lai, Brian and Dan Reiter. 2000. "Democracy, political similarity, and international alliances, 1816-1992." *Journal of Conflict Resolution* 44(2):203–227

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.

Levy, Jack S. 1981. “Alliance formation and war behavior: An analysis of the great powers, 1495-1975.” *Journal of Conflict Resolution* pp. 581–613.

Lewis, Jeffrey B and Kenneth A Schultz. 2003. “Revealing preferences: Empirical estimation of a crisis bargaining game with incomplete information.” *Political Analysis* 11(4):345–367.

Lovie, Sandy and Pat Lovie. 2010. “Commentary: Charles Spearman and correlation: a commentary on ’The proof and measurement of association between two things’.” *International journal of epidemiology* 39(5):1151–1153.

Miller, Delbert C and Neil J Salkind. 2002. *Handbook of research design and social measurement*. Sage.

Morrow, James D. 1991. “Alliances and asymmetry: An alternative to the capability aggregation model of alliances.” *American Journal of Political Science* pp. 904–933.

Ordeshook, Peter C. 1986. *Game theory and political theory: An introduction*. Cambridge University Press.

Organski, Abramo FK and Jacek Kugler. 1981. *The war ledger*. University of Chicago Press.

Poast, Paul. 2010. “(Mis) using dyadic data to analyze multilateral events.” *Political Analysis* 18(4):403–425

Pontius Jr, Robert Gilmore and Marco Millones. 2011. “Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment.” *International Journal of*

Remote Sensing 32(15):4407–4429.

Poole, Keith T. 2005. *Spatial models of parliamentary voting*. Cambridge University Press.

Poole, Keith T and Howard Rosenthal. 1985. “A spatial model for legislative roll call analysis.” *American Journal of Political Science* pp. 357–384.

Ray, James Lee. 2005. “Constructing multivariate analyses (of dangerous dyads).” *Conflict Management and Peace Science* 22(4):277–292.

Riker, William H. 1962. *The theory of political coalitions*. Yale University Press.

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

Scott, William A. 1955. “Reliability of content analysis: The case of nominal scale coding.” *Public opinion quarterly* pp. 321–325.

Sheskin, David J. 2004. *Handbook of parametric and nonparametric statistical procedures. 3rd ed.* crc Press.

Signorino, Curtis S and Jeffrey M Ritter. 1999. “Tau-b or not tau-b: measuring the similarity of foreign policy positions.” *International Studies Quarterly* 43(1):115–144.

Small, Melvin and J David Singer. 1969. “Formal alliances, 1816-1965: An extension of the basic data.” Journal of Peace Research 6(3):257–282.

Stevens, Stanley Smith et al. 1946. “On the theory of scales of measurement.”.

Tufte, Edward R. 1969. “Improving data analysis in political science.” World Politics 21(4):641–654.

Uebersax, John S. 1987. “Diversity of decision-making models and the measurement of interrater agreement.” Psychological bulletin 101(1):140.

Velleman, Paul F. and Leland Wilkinson (1993). "Nominal, ordinal, interval, and ratio typologies are misleading". The American Statistician. 47 (1): 65–72.

Viera, Anthony J, Joanne M Garrett et al. 2005. “Understanding interobserver agreement: the kappa statistic.” Fam Med 37(5):360–363.

Wallace, Michael D. 1973. “Alliance Polarization, Cross-Cutting, and International War, 1815-1964 A Measurement Procedure and Some Preliminary Evidence.” Journal of Conflict Resolution 17(4):575–604.

Zorn, Christopher JW. 2001. “Generalized estimating equation models for correlated data: A review with applications.” American Journal of Political Science pp. 470–490.

The Supplementary File

I. S-score as a Modified Spearman's ρ (or R)

Signorino and Ritter (1999) argued that the issues they have identified in measuring policy similarity "cannot surmount these problems by replacing τb with some other correlation coefficient like Spearman's ρ." The S-score, however, is based on essentially the same concept as the original form of Spearman's ρ. The original suggestion by Spearman (called R) was to use the absolute distance metric for calculating the original form of ρ as Signorino and Ritter (1999) used the metric for constructing the S-score as the numerator and the inverse of the most probable total sum of difference as the denominator (Spearman 1904, 86–88) (Kendall and Gibbons 1990, 9).[48]

Spearman's ρ treats the distances between neighboring ranks as the same (an interval), which explains the use of "rank" instead of the original observations. To use Spearman's ρ for ordered scales, one has to do a rank transformation of the original observations (Spearman 1904, 1906, 1910; Lovie 1995). Since Signorino and Ritter (1999) assumed the scoring rule as "setting the intervals to the rank values of alliance types", their scoring rule is equivalent to the rank transformation.[49] In addition, whether the rank order is (1,2,3,4) or (3,2,1,0),[50] the results are the same because the distance between ranks are the same regardless of the orderings. Thus, calculating the S-score begins by handling the data in exactly the same way as Spearman's ρ rank transformation.

The main difference between the S-score and Spearman's ρ is the denominators that normalize the scores. Signorino and Ritter (1999) used the maximum distance for the denominator because the maximum possible distance among the ranks are known in alliance portfolios or UN votes. In S-score for the alliance portfolio, the maximum possible total sum for distance is 3N, where 3 is the maximum distance

[48] The background history and motivation for developing the Spearman's ρ can be found in (Lovie and Lovie 2010; Perks 2010)

[49] Even though the S-score's scoring rule is equivalent to rank transformation, this does not mean that the S-score's alliance type numbers are ranks because (Signorino and Ritter 1999) impose "intervals" to them.

[50] The former is (Bueno de Mesquita 1975)'s ordering, and the latter is (Signorino and Ritter 1999)'s ordering

between alliance types and N is the number of states in the alliance portfoio. In general rankings, however, the number of ranks and the maximum possible total sum of the difference in ranks are usually not fixed. The number of ranks depends on the number of subjects. If ties are allowed, then the rank numbers are no longer consecutive. For example, suppose there are five subjects to be ranked. If three subjects tie in second place, the rankings in order are 1, (2,2,2), 5. If three subjects tie in third place, the rankings in order are 1, 2, (3,3,3).[51] Because the number of ties affects the numbering of ranks, the normalizing denominator of Spearman's ρ is different from the S-score.[52]

Furthermore, Spearman had to pursue different versions of numerators and denominators for theoretical needs (Kendall and Gibbons 1990, 9). [53] Regardless of the versions of Spearman's ρ, the essential idea is the same: the measure takes advantage of the ”distance” between ranks and normalize it. Moreover, the fact that the denominators were changed due to necessity by Spearman himself shows that S-score is also a modified version of ρ.

Therefore, counter to Signorino and Ritter (1999, 126)'s claim that they have developed a *new* measure of foreign policy similarity, the S-score is actually a modified version of Spearman's ρ. Thus, if Spearman's ρ cannot surmount the issues Signorino and Ritter (1999) raised, neither can the S-score.

II. The Means and Standard Deviations of the Measurements

The following table provides the information about the means and standard deviations of the measurements.

Note that ϕ global is strikingly different from Tau-b global because over 70% of the global data is missing in Eugene. However, the regional values of ϕ and tau-b are very similar. In addition, Leeds (2003) dataset, no missing data for all measurements, show that the means and standard deviations of tau-b and ϕ are very similar (69,836 observations in total). In

[51] The standard practice to deal with tied ranks is to take the average of tied ranks (Kendall and Gibbons 1990, Ch. 3). In that case, my example of 1,2,(3,3,3) would be 1,2,(4,4,4).
[52] For calculating Spearman's ρ with tied ranks, see (Kendall and Gibbons 1990, Ch.3)
[53] In the end, Spearman used “the sum of the distance squares metric” for the numerator ”the total sum of the square of the reverse of the other's ranking order (the maximum possible total square sum for distance)” as the denominator (Kendall and Gibbons 1990, p.8–9). For a detailed history of the evolution of the Spearman's ρ, see (Lovie 1995).

particular, the means of tau-b and ϕ are 0.0407612 and 0.041473, respectively. The standard deviations of tau-b and ϕ are 0.2799742 and 0.2821854, respectively.

	Mean	Std
Tau-b regional	-0.0595051	0.3091863
ϕ regional	-0.0602199	0.3143053
Tau-b global	0.021138	0.252495
ϕ global	0.0053797	0.2511745
S-score regional (unweighted)	0.443188	0.5390787
H-score regional (unweighted)	0.4165609	0.5543867
I-score regional (unweighted)	0.4109345	0.5544225
S-score global (unweighted)	0.7590874	0.1889808
H-score global (unweighted)	0.7222362	0.2494753
I-score global (unweighted)	0.7196243	0.2488629
S-score regional (weighted)	0.5206259	0.4535673
H-score regional (weighted)	0.4864987	0.4802634
I-score regional (weighted)	0.4766526	0.4827776
S-score global (weighted)	0.6807586	0.2894388
H-score global (weighted)	0.6575959	0.3074089
I-score global (weighted)	0.6513817	0.3097721

Table 1: Means and Standard Deviations of the measurements

III. The calculation of tau-b for Table 4 in the manuscript

The equation of tau-b for a tied-ranking order is

$$t_b = \frac{S}{\sqrt{[\frac{1}{2}n(n-1) - U]}\sqrt{[\frac{1}{2}n(n-1) - V]}}$$

Where

$$U = \frac{1}{2}\Sigma u(u-1)$$

And

$$V = \frac{1}{2}\Sigma v(v-1)$$

U is for ties in one ranking and V is in the other. S is the difference between concordant pairs and discordant pairs and n is the number of item.

To calculate tau-b, we need to construct concordant and discordant pairs between all possible combinations of the items. That is, we need $\binom{10}{2} = 45$ pairs of rankings for *i* and *j*.

The Table 4 in the main article is as follows

	i	*j*	s	Entry	$X_i = Y_i$
i	3	0	0.2	b	-1
j	0	3	0.2	c	-1
A	0	0	0	d	1
B	3	2	0.067	a	-1
C	1	2	0.067	a	-1
D	2	3	0.067	a	-1
E	1	2	0.067	a	-1
F	1	2	0.067	a	-1
G	1	0	0.067	b	-1
H	3	3	0	a	1

S-score	0.2	$s = \frac{2\lvert x_i - y_i \rvert}{30}$
H-score	0.4	$S = 1 - \Sigma s$
I-score	-0.6	Entry: as in Table 2.
ϕ-score	0.22	$X_i = Y_i \rightarrow 1$
Tau-b	0.12	$X_i \neq Y_i \rightarrow -1$

Table 4: A Concrete Example from the main article

The difference between concordant and discordant pairs is 4.
For i, the number of ties of 0, 1, 3 is 2, 4 and 3, respectively. Thus,
$U = \frac{2 \cdot 1 + 4 \cdot 3 + 3 \cdot 2}{2} = 10$.
For j, the number of ties of 0, 2, 3 is 3, 4 and 3, respectively. Thus,
$V = \frac{3 \cdot 2 + 4 \cdot 3 + 3 \cdot 2}{2} = 12$.

$$t_b = \frac{4}{\sqrt{[\frac{1}{2}10 \cdot 9 - 10]}\sqrt{[\frac{1}{2}10 \cdot 9 - 12]}} = 0.1177$$

IV. The replication of Leeds (2003)

The following table shows the replication of the ”Base Model Coefficient” in Leeds (2003, Table 1). Model (1) is the base model for this replication because it does not include any similarity measure. Model (2) represents the replication of Leeds (2003); it shows identical results as Leeds (2003, Table 1). Model (3) uses the updated S-score, (4) uses the H-score (global). Except the original model in Leeds (2003), replicated in Model (2), the Joint Democracy variable in all other models reduces dispute initiation at the 5% level of statistical significance.

Leeds (2003) tried to reconcile her result with existing studies by dividing the data into two groups: 1816-1914 and 1914-1944. For the sample of 1914-1944, Leeds (2003) showed that the Joint Democracy reduces dispute initiation. However, if we use an updated S-score or new similarity measures, without dividing the dataset, the results are consistent with the existing literature.

Table 2 is the base model. Table 3 is the replication of ”Effects of Outside Allies Coefficient” in Leeds (2003, Table 1), the core results of the study.

Model (1) is again the base model for this replication as before. Model (2) is the replication of Leeds (2003), which exactly replicates the results in Leeds (2003, Table 1).

In all models other than the replicated Model (2), the Joint Democracy variable significantly reduces dispute initiation at less than the 5% level of statistical significance. However, aside from the original model, (2), in no other model is the defense pact variable, one of the core independent variables, significant at any conventional level.

Regarding similarity measures, the new S-score and the H-score are all quite comparable. The H-score meets the 5% level of statistical significance, whereas the new S-score is not statistically significant at any conventional level. Tau-b and φ are also quite similar.

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Joint Democracy	-0.503**	-0.353	-0.499**	-0.522**	-0.517**	-0.500**	-0.500**
	(0.239)	(0.228)	(0.236)	(0.235)	(0.235)	(0.238)	(0.238)
Contiguity	1.000***	1.112***	1.077***	1.114***	1.104***	1.010***	1.010***
	(0.152)	(0.151)	(0.155)	(0.155)	(0.156)	(0.149)	(0.149)
PowerofPotentialChallengersinRelationto	0.792***	0.728***	0.771***	0.760***	0.766***	0.790***	0.790***
Potential Target	(0.150)	(0.148)	(0.148)	(0.147)	(0.148)	(0.149)	(0.149)
Shared Alliance Commitment	-0.417*	-0.297	-0.352	-0.361	-0.332	-0.246	-0.241
	(0.226)	(0.227)	(0.223)	(0.222)	(0.222)	(0.332)	(0.330)
S-score (Old) S-		-0.940***					
		(0.121)					
score (New)			-0.330**				
			(0.153)				
I-score				-0.361***			
				(0.134)			
H-score					-0.362***		
					(0.129)		
Tau-b						-0.264	
						(0.306)	
ϕ							-0.271
							(0.302)
Constant	-5.191***	-4.666***	-5.032***	-5.051***	-5.048***	-5.195***	-5.195***
	(0.130)	(0.132)	(0.146)	(0.136)	(0.137)	(0.130)	(0.130)
Observations	69,730	69,730	69,730	69,730	69,730	69,730	69,730
Number of dyad	1,364	1,364	1,364	1,364	1,364	1,364	1,364

Robust standard errors in parentheses
**** p<0.01, ** p<0.05, * p<0.1*

Table 2: The Replication of "*Base Model Coefficient*" of Table 1 in (Leeds 2003)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Joint Democracy	-0.447**	-0.323	-0.443**	-0.457**	-0.453**	-0.447**	-0.447**
	(0.225)	(0.215)	(0.222)	(0.222)	(0.222)	(0.224)	(0.224)
Contiguity	1.124***	1.232***	1.178***	1.207***	1.201***	1.131***	1.131***
	(0.144)	(0.144)	(0.146)	(0.145)	(0.147)	(0.142)	(0.142)
Power of Potential Challengers in Relation to Potential Target	0.582***	0.539***	0.569***	0.560***	0.563***	0.580***	0.580***
	(0.147)	(0.146)	(0.146)	(0.145)	(0.146)	(0.147)	(0.147)
Shared Alliance Commitment	-0.494**	-0.292	-0.429*	-0.433*	-0.406*	-0.384	-0.377
	(0.234)	(0.233)	(0.233)	(0.232)	(0.232)	(0.321)	(0.319)
S-score (Old) S-		-0.916***					
		(0.122)					
score (New)			-0.260				
			(0.165)				
I-score				-0.295**			
				(0.139)			
H-score					-0.298**		
					(0.137)		
Tau-b						-0.174	
						(0.290)	
ϕ							-0.183
							(0.285)
Potential Target Has Defensive Ally	-0.184	-0.331**	-0.210	-0.212	-0.214	-0.182	-0.182
	(0.142)	(0.139)	(0.143)	(0.141)	(0.142)	(0.142)	(0.142)
Potential Challenger Has Offensive Ally	0.483***	0.390***	0.445***	0.443***	0.443***	0.480***	0.480***
	(0.134)	(0.127)	(0.137)	(0.134)	(0.135)	(0.134)	(0.134)
PotentialChallengerHasRelevantNeutralityPact	0.554***	0.461***	0.561***	0.551***	0.551***	0.550***	0.550***
	(0.115)	(0.116)	(0.114)	(0.115)	(0.114)	(0.116)	(0.116)
Constant	-5.225***	-4.666***	-5.090***	-5.097***	-5.093***	-5.227***	-5.227***
	(0.139)	(0.151)	(0.168)	(0.153)	(0.155)	(0.140)	(0.140)
Observations	69,730	69,730	69,730	69,730	69,730	69,730	69,730
Number of dyad	1,364	1,364	1,364	1,364	1,364	1,364	1,364

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1*

Table 3: TheReplication of "*Effectsof Outside AlliesCoefficient*"of Table 1in(Leeds2003)

References

Bueno de Mesquita, Bruce. 1975. "Measuring systemic polarity." *Journal of Conflict Resolution* 19(2):187–216.

Kendall, Maurice George and Jean Dickinson Gibbons. 1990. *Rank correlation methods*. Edward Arnold. Oxford University Press.

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.

Lovie, Alexander D. 1995. "Who discovered Spearman's rank correlation?" *British Journal of Mathematical and Statistical Psychology* 48(2):255–269.

Lovie, Sandy and Pat Lovie. 2010. "Commentary: Charles Spearman and correlation: a commentary on 'The proof and measurement of association between two things'." *International journal of epidemiology* 39(5):1151–1153.

Perks, Julie. 2010. "Commentary:The next trick is impossible." *International journal of epidemiology* 39(5):1153–1155.

Signorino, Curtis S and Jeffrey M Ritter. 1999. "Tau-b or not tau-b: measuring the similarity of foreign policy positions." *International Studies Quarterly* 43(1):115–144.

Spearman, Charles. 1904. “The proof and measurement of association between two things.” *The American journal of psychology* 15(1):72–101.

Spearman, Charles. 1906. “’Footrule’ for measuring correlation.” *British Journal of Psychology* 2(1):89–108.

Spearman, Charles. 1910. “Correlation calculated from faulty data.” *British journal of psychology* 3(3):271–295.