Advance Access publication July 16, 2011

Political Analysis (2011) 19:287–305
doi:10.1093/pan/mpr023

Choice or Circumstance? Adjusting Measures of Foreign Policy Similarity for Chance Agreement

Frank M. Häge
Department of Politics and Public Administration, University of Limerick, Limerick, Ireland
e-mail: frank.haege@ul.ie

The similarity of states' foreign policy positions is a standard variable in the dyadic analysis of international relations. Recent studies routinely rely on Signorino and Ritter's (1999, Tau-b or not tau-b: Measuring the similarity of foreign policy positions. *International Studies Quarterly* 43:115–44) *S* to assess the similarity of foreign policy ties. However, *S* neglects two fundamental characteristics of the international state system: foreign policy ties are relatively rare and individual states differ in their innate propensity to form such ties. I propose two chance-corrected agreement indices, Scott's (1955, Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly* 19:321–5) π and Cohen's (1960, A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46) κ, as viable alternatives. Both indices adjust the dyadic similarity score for a large number of common absent ties. Cohen's κ also takes into account differences in individual dyad members' total number of ties. The resulting similarity scores have stronger face validity than *S*. A comparison of their empirical distributions and a replication of Gartzke's (2007, The capitalist peace. *American Journal of Political Science* 51:166–91) study of the 'Capitalist Peace' indicate that the different types of measures are not substitutable.

1 Introduction

The similarity of states' foreign policy positions is a standard variable in the quantitative, dyadic analysis of international relations. The variable is supposed to capture the extent to which pairs of states have shared or opposing interests. Explicitly or implicitly, the degree of similar or opposing state interests forms part of most explanations for international cooperation and conflict. For example, similar state interests are hypothesized to foster bilateral trade (Morrow, Siverson, and Tabares 1998; Kastner 2007), to increase the chances of receiving military and development aid (Neumayer 2003; Derouen and Heo 2004), to improve the effective functioning of international institutions (Stone 2004), to reduce the incentives to harbor foreign terrorist groups (Bapat 2007), and, of course, to decrease the risk of conflict and militarized disputes (Bearce, Flanagan, and Floros 2006; Long and Leeds 2006; Gartzke 2007; Braumoeller 2008).

Yet despite the importance of this variable, the measurement of foreign policy similarity has received little attention. Bueno de Mesquita (1975) originally proposed Kendall's (1938) rank-order correlation coefficient τ_b as a measure of similarity. According to this measure, the foreign policy ties of two states are maximally similar if their rankings exhibit perfect covariation. Signorino and Ritter (1999) objected to the use of τ_b on conceptual grounds. They argue that τ_b does not indicate the extent to which two states share the same types of foreign policy ties to other states, but only the extent to which the two states rank their foreign policy ties to other states in a similar manner (Signorino and Ritter 1999, 121). Signorino and Ritter (1999) propose *S* as an alternative measure. According to this measure, the foreign policy tie profiles of two states are maximally similar if they match exactly, regardless of whether or not the strength of foreign policy ties covaries. Signorino and Ritter's *S* has since become the prevailing measure of foreign policy positions in the statistical analyses of international relations.[1] Despite its growing popularity, few studies have subsequently examined the properties of *S*. Although Bennett and Rupert (2003) and Sweeney and Keshk (2005) have pointed to some empirical and conceptual problems of *S*, they have not suggested feasible alternatives.

[1] A search in the Social Science Citation Index for articles citing Signorino and Ritter (1999) returns 126 matches (http://isiwebofknowledge.com [accessed April 26, 2011]). A similar search in Google Scholar returns 273 hits (http://scholar.google.com [accessed April 26, 2011]). Although *S* might not have completely replaced τ_b, I am not aware of any recent study that relies exclusively on τ_b, without reporting results with *S* as well. Supplementary materials for this article are available on the *Political Analysis* Web site.

In this article, I discuss the application of chance-corrected agreement indices to assess the similarity of states' foreign policy positions. Even though these measures have been developed in a different context for different research applications, their conceptual properties make them uniquely suited for measuring the similarity of foreign policy positions in the study of international relations. The inquiry is motivated by the observation that *S* often yields implausible similarity scores. The lack of face validity of *S* is illustrated in the left panel of Fig. 1. The figure indicates similarity scores of the United Kingdom with the other four permanent members of the United Nations Security Council during the Cold War. In line with most existing research, the reported *S* values are based on data about dyad members' alliance ties with all other states in the international system. The assumption underlying the use of these data is that any similarity in alliance commitments is a result of similar foreign policy positions (Altfeld and Mesquita 1979, 116). We know that the U.K.'s security interests during the Cold War were relatively close to those of France and the United States. At the same time, the United Kingdom had very different interests from those of China and the Soviet Union. The *S* values for the U.K.-France dyad and the U.K.-U.S. dyad are roughly in line with the historical record. However, the U.K.-Soviet Union and U.K.-China dyads show *S* values that are too high in comparison. During the entire period, the U.K.'s *S* score with the Soviet Union is very similar and sometimes even higher than its *S* score with the United States. The *S* values for the U.K.-China dyad are even more implausible. They indicate that, during the entire Cold War period, the interests of the United Kingdom were considerably more similar to those of China than to those of the United States.

In the remainder of this article, I show that the lack of face validity of *S* is a result of the measure's way of standardizing the extent of dissimilarity of states' foreign policy tie profiles. At its core, *S* measures the dissimilarity of states' tie profiles and adjusts it for the theoretically possible maximum dissimilarity. Features of the observed empirical distributions of individual dyad members' foreign policy ties are not taken into account. In substantive terms, the distribution-independent standardization in the calculation of *S* implies that the measure neglects two fundamental aspects of the international state system: the low density of foreign policy ties in the system and the innate differences of individual states to form such ties. In contrast, chance-corrected agreement indices offer distribution-dependent ways of standardizing the extent of dissimilarity. Scott's (1955) π and Cohen's (1960) κ adjust the observed dissimilarity of tie profiles for a generally low propensity of dyad members to form foreign policy ties. In addition, κ takes into account that individual dyad members may differ in their propensity to form ties.

Before discussing the calculation and advantages of these indices in more detail, it is worth having a preliminary look at the resulting empirical differences in the similarity values of the different measures. The middle and right panel of Fig. 1 show the U.K.'s similarity scores based on π and κ, respectively. The similarity scores for the U.K.-France and the U.K.-U.S. dyad remain positive and relatively large when the

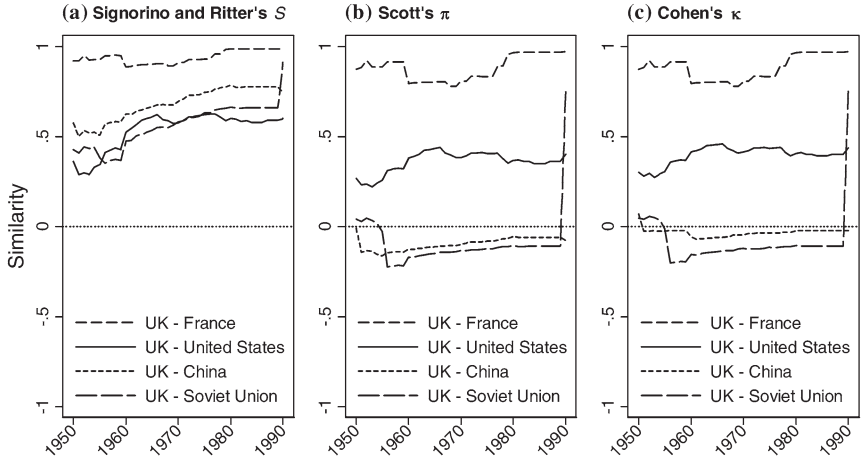

Fig. 1 Similarity values of dyads involving the United Kingdom (1950–1990). The figure compares similarity values of dyads involving the United Kingdom and other permanent members of the U.N. Security Council during the Cold War. Panel (a) shows similarity values generated by Signorino and Ritter's *S* with a squared distance metric, panel (b) shows values generated by Scott's π, and panel (c) shows values generated by Cohen's κ. The measures are based on alliance data for all members of the international state system (Correlates of War Project 2003, 2005).

agreement indices are applied instead of *S*. However, the scores for the U.K.-China and the U.K.-Soviet Union dyad are now much lower and consistently negative for most of the time period examined. Although the differences between π and κ are generally small, the significantly lower π score of the U.K.-China dyad is noteworthy. Keeping their minor differences in mind, we can conclude for the moment that the two chance-corrected agreement indices produce similarity scores that are clearly more in line with the conventional wisdom about states' foreign policy positions during the Cold War than *S*.

In the next section, I present a brief review of how similarity scores are calculated according to *S*. Then I describe in more detail the two weaknesses of *S* that result in implausible similarity scores. Having identified the problems affecting *S*, I propose Cohen's κ and Scott's π as two useful alternatives and describe their computation. For conceptual clarity and ease of exposition, my discussion of the limitations of *S* and the computation of π and κ relies on data about binary foreign policy ties. As most current applications use valued tie data to measure the similarity of foreign policy positions, I subsequently describe extensions of π and κ to assess the similarity of ties whose strength is measured on a quantitative scale.[2] Empirical comparisons of those measures show that the distribution of *S* strongly differs from the distributions of the two chance-corrected indices. Thus, the two types of measures are clearly not interchangeable. This conclusion is also confirmed by a replication of Gartzke's (2007) study of the "Capitalist Peace." The replication results demonstrate that the two types of similarity measures can lead to substantially different statistical inferences. Although the two chance-corrected indices yield similar values and replication results in these examples, the use of one or the other implies very different assumptions about the data generation process. Thus, the decision about which chance-corrected index to apply should be guided by theoretical considerations.

2 Measuring the Similarity of Foreign Policy Positions with *S*

Foreign policy positions of states are hard to observe directly. One way to measure them is to rely on an indicator of observable behavior that "reveals" the preferences responsible for generating that behavior. Traditionally, data on alliance portfolios have been used to assess the similarity of foreign policy positions (Altfeld and Mesquita 1979). For ease of exposition, I follow this convention in the conceptual discussion and comparison of similarity measures.[3] The assumption underlying the use of alliance data is that similar alliance portfolios are the result of similar foreign policy positions. A state's alliance portfolio can be represented in the form of a vector, where the individual entries of the vector indicate the existence and the strength of its alliance commitments with other states in the international system. The strength of alliance commitments can range from "no commitment," "entente," "neutrality or nonaggression pact," to "defense pact."[4] All similarity measures aim to assess the dissimilarity of the alliance commitment vectors of the two dyad members and then convert it into a similarity score. The measures even calculate the dissimilarity of the two vectors in exactly the same way. Both *S* and the chance-corrected agreement indices appraise dissimilarity through a simple distance function. The subsequent conversion of the dissimilarity into a similarity score is also the same. The measures only differ in the way in which dissimilarity values are standardized. As mentioned earlier, these differences in the standardization result in crucial differences in similarity scores.

[2] I borrow the distinction between binary and valued data from social network analysis. The former type of data indicates only the presence or absence of ties, whereas the latter also indicates the strength of ties (Scott 2000, 47). The version of κ for valued or quantitative data is also known as weighted κ (Cohen 1968), a_F (Krippendorff 1970), chance-corrected identity coefficient (Zegers 1986), concordance correlation coefficient (Lin 1989), and fixed marginal agreement coefficient (Fay 2005). The version of π for quantitative data is also known as a_R (Krippendorff 1970) and random marginal agreement coefficient (Fay 2005). All similarity measures discussed in this article are available for download from http://www.frankhaege.eu. This collection includes measures based on binary as well as valued ties for all state system members (Correlates of War Project 2005), calculated from alliance (1816–2000) and U.N. voting data (1946–2004). The "rmac" package (Kirk 2010) in R (R Development Core Team 2011) implements the computation of the interrater agreement indices as described by Fay (2005). A complete replication archive for the analyses conducted in this article is available from the Political Analysis dataverse at http://hdl.handle.net/1902.1/16052.

[3] In principle, similarity measures can be calculated with any relational data set (Sweeney and Keshk 2005). Besides alliance data, the use of data on voting in the U.N. General Assembly has been popular (Gartzke 1998). The empirical comparison in Section 5 involves measures based on U.N. voting data as well.

[4] The vectors also include an entry for each of the dyad members. Relationships of states to themselves are coded as defence pacts and therefore receive the maximum scale value (Bueno de Mesquita 1975, 195).

Before turning to a discussion of the chance-corrected agreement indices, I review the computation of *S* and investigate the reasons for its lack of face validity. Equation (1) presents a simplified version of the formula for the calculation of *S*:

$$S = 1 - 2 \times \frac{\sum |X_i - Y_i|}{\sum d_{\max}}. \quad (1)$$

The more general formula given by Signorino and Ritter (1999, 127) does not specify a specific distance metric and allows for the incorporation and differential weighting of additional types of foreign policy ties. In practice, most existing research has relied on the absolute value distance metric $|X_i - Y_i|$ and a single data source to calculate *S* values.[5] The more general formula for *S* also allows weighting foreign policy ties by countries' importance. Indeed, Signorino and Ritter (1999, 133) suggest weighting ties by countries' material capabilities (Correlates of War Project 2005) to deal with the problem of the preponderance of absent alliance ties. However, this procedure is problematic. The distribution of material capabilities of states is extremely skewed. Capability-weighting means that a lot of information about most countries is effectively discarded from the sample (e.g., the five largest powers in 1985 contribute more than 50% to the calculation of dissimilarity values in that year).

The effective restriction of the sample to a few very powerful states leads only to more plausible *S* scores if those states have a higher propensity to establish foreign policy ties than the excluded, less powerful ones. Such a relationship will then result in generally less and more meaningful shared absences of foreign policy ties. Although such a positive association between material capabilities and the total number of foreign policy ties indeed exists, it is far from perfect. Thus, weighting is at best a second best solution. Chance-corrected agreement indices provide a solution that does not rely on an additional data source with all its potential for introducing further measurement error. Also, a more fundamental objection is that weighting is essentially a sampling decision and should be made independent of measurement issues. If weighting ties is deemed desirable to assess the similarity of foreign policy positions, then any measure of similarity should be calculated on weighted data, including Scott's π and Cohen's κ.

Relying on equation (1), Fig. 2 illustrates the calculation of *S* with a hypothetical example. The data consist of binary alliance ties indicating the presence (1) or absence (0) of any alliance commitment between dyad members *X* and *Y* and the other states *A* to *H* in the international system. Relying on binary data makes the exposition easier and allows us to distinguish qualitative properties of the different measures from properties of the distance metric. Unless stated otherwise, all discussed issues apply analogously to valued tie data as well. Regarding the entries in panel (a) of Fig. 2, the first row of the matrix indicates that state *X* has an alliance commitment to itself, whereas state *Y* does not have an alliance commitment to state *X*. The second row indicates the converse situation. State *Y* has an alliance commitment to itself but not to state *X*. The other rows provide information about the two dyad members' alliance commitments to the remaining states in the international system. The two tie profiles are similar in that both states share an alliance commitment to state *E* and both do not have alliance commitments with states *C*, *D*, *G*, and *H*. However, the two tie profiles are dissimilar in that only *X* has an alliance commitment to *B* and only *Y* has an alliance commitment to *A* and *F*.

We can compute the *S* score of the *X*-*Y* dyad directly from the matrix given in panel (a) of Fig. 2. First, we calculate the absolute distances $|X_i - Y_i|$ between the entries in columns *X* and *Y*. In a second step, we sum the absolute distances across rows, which yields the observed dissimilarity ($D_o = 5$) of the two tie profiles. The observed dissimilarity is then standardized by dividing it by the maximum possible dissimilarity ($D_{\max}$). In the case of binary data, the maximum possible dissimilarity of individual alliance ties ($d_{\max}$) is 1, so the maximum possible dissimilarity is 1 times the number of countries *n*, which results in $D_{\max} = nd_{\max} = n = 10$. The resulting measure is a proportion that can take values between 0 and 1. In this case, the proportion of dissimilar ties is 0.5 ($P_d = D_o/D_{\max} = 5/10 = 0.5$).

We can derive the proportion of dissimilarity value more easily from a contingency table. The contingency table view is useful because it provides a straightforward summary of the main features of the

[5]The squared distance metric would be a prominent alternative.

(a) Column vectors

	X	Y	$\lvert X_i\text{-}Y_i\rvert$
X	1	0	1
Y	0	1	1
A	0	1	1
B	1	0	1
C	0	0	0
D	0	0	0
E	1	1	0
F	0	1	1
G	0	0	0
H	0	0	0
		Σ	**5**

Calculation

$D_o = \lvert X_i - Y_i\rvert$	5
$D_{max} = \Sigma d_{max}$	10
$P_d = D_o/D_{max}$	.50
$S = 1 - 2P_d$	.00

(b) Contingency table

Absolute frequencies

		Y = 0	Y = 1	
X	0	4	3	7
	1	2	1	3
		6	4	10

Relative frequencies

		Y = 0	Y = 1	
X	0	.40	.30	.70
	1	.20	.10	.30
		.60	.40	1

Calculation

$D_o = b + c$	5
$D_{max} = nd_{max}$	10
$P_d = D_o/D_{max}$	.50
$S = 1 - 2P_d$	.00

Fig. 2 Two ways of calculating *S*. The figure illustrates the calculation of *S* from hypothetical, binary alliance data of dyad members *X* and *Y*. Panel (a) demonstrates the calculation of *S* directly from the two column vectors representing alliance portfolios. Panel (b) demonstrates the calculation of *S* from the data contained in the contingency table of the two alliance portfolios. $N = 10$ countries; 0 denotes the absence and 1 denotes the presence of an alliance commitment.

data that are relevant for assessing the dissimilarity of two tie profiles. The entries in the diagonal cells of the contingency table with absolute frequencies ($a = 4$ and $d = 1$) in panel (b) of Fig. 2 indicate the number of similar alliance ties and the entries in the off-diagonal cells ($b = 3$ and $c = 2$) indicate the number of dissimilar alliance ties. The observed total dissimilarity between tie profiles is equal to the total number of dissimilar alliance ties and can be derived by simply adding up the entries in the off-diagonal cells of the table ($D_o = 3 + 2 = 5$). Dividing the total number of dissimilar alliance ties by the total number of countries yields the proportion of dissimilarity ($P_d = D_o/D_{\max} = 5/10 = 0.5$).

When the contingency table indicates relative rather than absolute frequencies, the proportion of dissimilarity can be computed even more directly by just adding up the relative frequencies in the off-diagonal cells of the table.[6] In this case, no further division by the total number of countries is required. No matter how the proportion of dissimilarity is derived, it is subsequently multiplied by 2 and subtracted from 1 to transform it from a dissimilarity measure with a theoretical range between 0 and 1, where 1 indicates complete dissimilarity, to a similarity measure with a theoretical range between −1 and 1, where 1 indicates complete similarity. In the example, this linear transformation results in an *S* value of zero ($S = 1 - 2[0.5] = 0$).

The example shows that *S* scores are a direct function of the proportion of dissimilarity. The proportion of dissimilarity has the same value regardless of whether dissimilar ties are distributed equally across all off-diagonal cells or concentrated in one of those cells. The proportion is also not affected by the distribution of similar ties across the diagonal cells of the contingency table. This insensitivity of *S* to the

[6]Because of this simplicity, I make extensive use of contingency tables presenting relative frequencies or proportions in the remainder of this article.

distribution of similar and dissimilar ties across their respective cells in the contingency table is problematic.[7] As Fig. 3 demonstrates, the same proportion of dissimilarity can mean very different things, mainly as a result of how dyad members' marginal distributions constrain the way similar and dissimilar ties are distributed over their respective cells.[8]

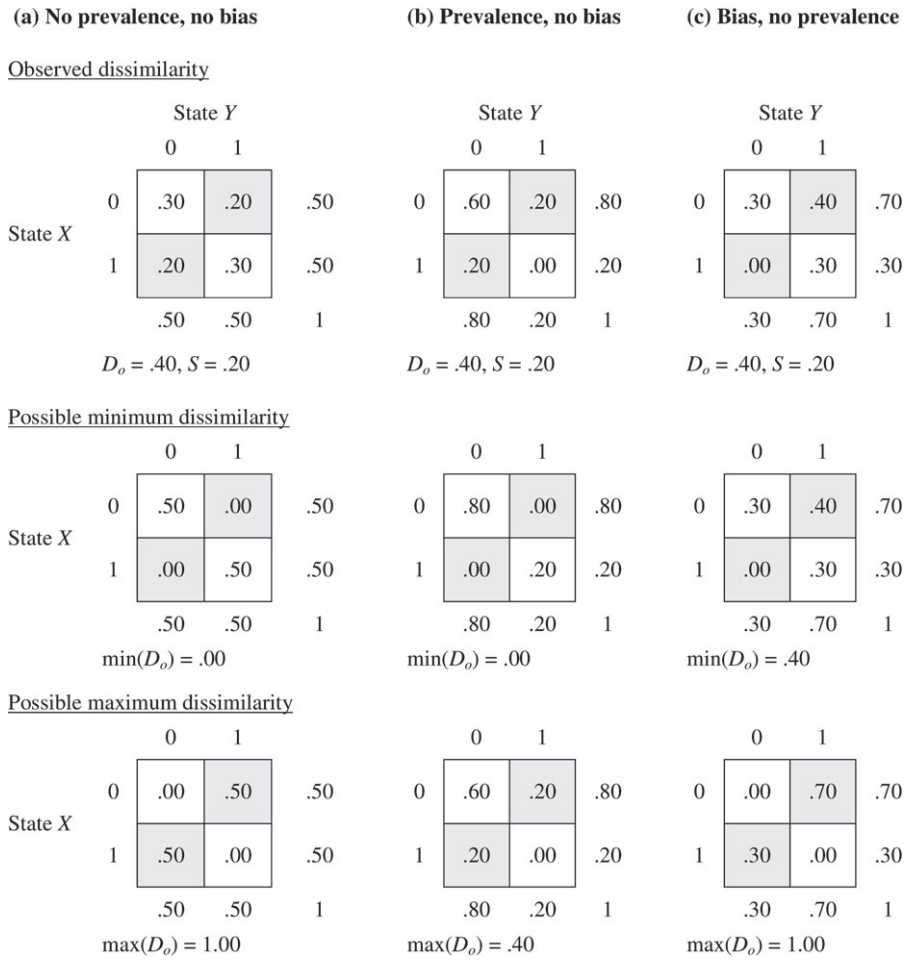

Fig. 3 Unbalanced marginal distributions do not affect *S*. ''Prevalence'' stands for symmetrically unbalanced marginal distributions; ''bias'' for asymmetrically unbalanced marginal distributions. All cell entries are proportions (relative frequencies). D_o denotes the proportion of dissimilarity and is the sum of the entries in the off-diagonal cells of the contingency tables. Multiplying D_o by 2 and subtracting it from 1 creates Signorino and Ritter's *S*. Minimum dissimilarity is the smallest proportion of ties that has to be in the off-diagonal cells, given the observed distribution of the marginals. Maximum dissimilarity is the maximum proportion of ties that could be in the off-diagonal cells, given the observed distribution of marginals.

[7]According to Signorino and Ritter (1999, 121–3), the main advantage of *S* over association measures like is exactly this insensitivity to the lack or form of covariation. However, inferring the similarity of foreign policy positions by comparing dyad members' behavior is only possible by assessing the degree to which dyad members' behavior varies in similar ways. Covariation between variables is generally accepted as one of the main conditions for establishing causality (De Vaus 2001, 34; Kellstedt and Whitten 2008, 48). Thus, if two tie profiles do not covary or covary in a negative way, then they are clearly not causally related to similar foreign policy positions. Also in this sense, Scott's π and Cohen's κ are improvements. Cohen's κ is actually a chance-corrected measure of association (Zegers 1986), and Scott's π will never indicate a positive similarity value in the absence of a positive association between the tie profiles (Fay 2005, 175). From this point of view, the problem of is not its reliance on covariation, but its lack of chance correction.

[8]The issues discussed here have long been identified in the literature on assessing interrater agreement. However, in that context, they have usually been interpreted as problems of Cohen's κ rather than problems of the proportion of dissimilarity (Cicchetti and Feinstein 1990; Feinstein and Cicchetti 1990; Byrt, Bishop, and Carlin 1993; Lantz and Nebenzahl 1996; Sim and Wright 2005). For an exception, see Vach (2005).

In panel (a), the marginal distributions indicate that both dyad members have a 50% propensity to establish an alliance. In this case, the marginal distribution put no constraint on the empirically possible minimum and maximum dissimilarity value. As the lower two tables in panel (a) illustrate, the proportion of dissimilarity could take any value between 0% and 100%, so the observed proportion of dissimilarity of 40% provides a reasonable assessment of the two tie profiles' dissimilarity. The situation is different in panels (b) and (c). Panel (b) depicts the case in which both states have a low propensity of only 20% to establish an alliance. In other words, the marginal distributions are symmetrically unbalanced. As a result, non-alliance ties are generally more ''prevalent'' than alliance ties.[9] In this case, a proportion of dissimilarity of 40% is much more ''impressive.'' Given the marginal distributions, the lower two tables in panel (b) demonstrate that the proportion of dissimilarity can only vary between 0% and 40%. Thus, the observed proportion of dissimilarity is actually at its empirically possible maximum. The two alliance profiles could not be any more different in this situation. However, the proportion of dissimilarity and, by implication, *S* does not reflect this fact. Panel (c) presents the case in which the two dyad members differ strongly in their propensity to establish an alliance. In other words, they exhibit ''biased'' propensities. State *X* has a high propensity of 70%, but state *Y* has only a low propensity of 30%. This means the marginal are asymmetrically unbalanced. In this case, the proportion of dissimilarity of 40% is not very impressive. Given dyad members' marginal distributions, the lower two tables of panel (c) show that the proportion of dissimilarity can only take values between 40% and 100%. Thus, the observed proportion of dissimilarity is actually at its empirically possible minimum; it could not be any smaller. This information is also not reflected in the *S* score.

As discussed earlier, networks in international relations often exhibit low density. The establishment and maintenance of bilateral relationships between states are usually costly. As a result, these relationships are relatively rare and the absence of ties is much more common. Panel (b) of Fig. 3 illustrates that in such situations the unadjusted proportion of dissimilarity tends to indicate too little dissimilarity, resulting in *S* scores that seem too high. Although this prevalence of non-ties is widely accepted as a problem for measuring the similarity of foreign policy ties (Signorino and Ritter 1999, 124; Bennett and Rupert 2003, 372; Sweeney and Keshk 2005, 175), the differential biases of states as a source of implausible similarity values might be more controversial. Yet if foreign policy ties are costly to establish and maintain, then states are likely to differ systematically in their willingness and capability to bear such costs. For example, the United States are much more prepared and able to maintain an extensive net of alliance partners around the globe than Luxembourg. The degree to which states engage in alliance commitments is first and foremost driven by capability. If Luxembourg does not have the capability to come to the aid of Bolivia and Bolivia does not have the capability to come to the aid of Luxembourg in the case of a militarized conflict, then a defence pact between those two states seems unlikely. But even if a state is generally capable of projecting its force overseas, it might regard military alliances as an inadequate means to pursue its security interests and therefore consciously limit its engagement in these kinds of international arrangements. Neither the capability nor the general willingness to engage in military alliances reveals underlying foreign policy preferences. Foreign policy preferences are mainly reflected in the state's choice of alliance partners given a specific propensity to engage in such behavior, less so in the propensity to engage in such behavior itself. Panel (c) of Fig. 3 demonstrates that such differences in the propensity of states to form alliances will result in an unadjusted proportion of dissimilarity that indicates more dissimilarity than warranted, resulting in similarity scores of *S* that seem too low.

The independence of the propensity to form foreign policy ties from the choice of partner might vary with the costs involved in establishing and maintaining a tie. The assumption is certainly quite plausible in the case of alliance commitments but might be less justifiable for ''cheaper'' types of ties. A prime example of less costly relationships would be ties formed through identical or similar voting in the U.N. General Assembly (e.g., Gartzke 1998). In this case, the act of voting is equally costly, regardless of

[9]See especially Byrt, Bishop, and Carlin (1993) for the use of the terms ''prevalence'' and ''bias'' in the context of agreement indices. ''Prevalence'' refers to symmetrically unbalanced marginal distributions in a contingency table, and ''bias'' to asymmetrically unbalanced marginal distributions. In the case of alliance data, prevalence manifests itself in the preponderance of no-alliance ties and bias in a strongly disparate number of alliance partners of the two dyad members (e.g., the U.S.-Switzerland dyad post-World War II).

whether the country votes "Yes," "Abstain," or "No." The only cost a country might incur in these situations is directly related to which other countries it chooses to support or oppose through its vote. In this case, asymmetrically distributed marginals are mainly due to real differences in foreign policy positions. Fortunately, chance-corrected agreement indices can handle both situations. Cohen's κ corrects the proportion of dissimilarity for both prevalence and bias, but Scott's π only corrects it for the prevalence of a certain type of tie. The latter measure is therefore more appropriate in the case where foreign policy ties are cheap. In the next section, I describe the computation of both measures.

3 Chance Correction to Account for Prevalence and Bias

A certain proportion of dissimilar alliance ties is "harder" to achieve in the face of symmetrically unbalanced marginal distributions (i.e., prevalence) than in the face of balanced marginal distributions. In contrast, the same proportion of dissimilarity is "easier" to achieve in the face of asymmetrically unbalanced marginal distributions (i.e., bias) than in the face of balanced marginal distributions. Thus, the proportion of dissimilarity needs to be adjusted upward in the case of prevalence and downward in the case of bias. Chance-corrected agreement indices accomplish both these tasks. In general, these indices take the following form (e.g., Krippendorff 1970, 140):

$$\textit{Chance–corrected agreement} = 1-\frac{D_o}{D_e}.$$

D_o stands for the observed dissimilarity and D_e for the dissimilarity expected by chance. In the case of binary data, D_o is the sum of the proportions p_{ij} in the off-diagonal cells of the contingency table, where $i,j = 0, \ldots, k$ indicate the row and column numbers:

$$D_o = \sum_{i \neq j} p_{ij}. \tag{3}$$

D_e is calculated by multiplying the hypothesized marginal proportions of the two raters $m_{i\cdot}$ and $m_{j\cdot}$ for each off-diagonal category ij and by adding up the resulting products:

$$D_e = \sum_{i \neq j} m_{i\cdot} m_{\cdot j}. \tag{4}$$

The only difference in the calculation of chance-corrected agreement measures lies in the definition of the hypothesized marginal proportions $m_{i\cdot}$ and $m_{\cdot i}$ (Zwick 1988, 376). Even S can be reformulated as a chance-corrected agreement index. In this case, the hypothesized marginal proportions are $m_{i\cdot} = m_{\cdot j} = ½$ for both the highest and lowest rating category and zero otherwise. Plugging these values into equation (4), chance disagreement for S is calculated as follows:

$$D_e^S = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2}. \tag{5}$$

Inserting this result and the right-hand side of equation (3) into equation (2) yields S, expressed as a chance-corrected agreement index:

$$S = 1-\frac{\sum_{i \neq j} p_{ij}}{\frac{1}{2}}. \tag{6}$$

Equation (6) can easily be reformulated to $S = 1-2\sum_{i \neq j} p_{ij}$. We have seen earlier that the proportion of dissimilarity $\sum_{i \neq j} p_{ij}$ is the same as the standardized distance $\sum |X_i - Y_i| \sum d_{\max}$ in equation (1), thus the two formulas are equivalent. The main difference between S and chance-corrected agreement indices is that the "chance correction" of S is calculated independently of the observed marginal distributions of the two dyad members, whereas the chance corrections of π and κ take the form of the marginal

distributions into account. For *S*, the expected dissimilarity is always 50%, regardless of the prevalence of a certain tie or the differential biases of dyad members. In the case of binary data, the chance correction of *S* reduces to the assumption that all states are just as likely to form alliance ties as they are likely to form non-alliance ties.[10] The expected dissimilarity defines the zero value of chance-corrected coefficients (Krippendorff 2004, 416). Thus, whenever the actual marginal distributions deviate from the form of the marginal distributions assumed by *S*, the similarity values of *S* will be over- or understated.

Scott's π chance correction adjusts for prevalence but assumes that states do not exhibit any biases. In our context, the measure assumes that all states have a similar propensity to engage in military alliances; the "true" marginal distributions of the two dyad members are supposed to be homogenous (Zwick 1988, 367). However, unlike *S* which assumes an identical propensity of 0.5, π does not assume that states' common propensity to form alliance ties takes any specific value. Rather, dyad members' common propensity is estimated from the data in the contingency table by averaging the respective marginal proportions:

$$m_{i\cdot} = \frac{(p_{i\cdot} + p_{\cdot i})}{2} \text{ and } m_{\cdot j} = \frac{(p_{j\cdot} + p_{\cdot j})}{2}.$$

The estimated marginal proportions are then used to calculate the expected chance dissimilarity by plugging them into equation (4). Given the assumption of homogeneous marginal distributions, the chance correction of π takes the following form:

$$D_e^{\pi} = \sum_{i \neq j} \left(\frac{p_{i\cdot} + p_{\cdot i}}{2}\right)\left(\frac{p_{j\cdot} + p_{\cdot j}}{2}\right). \tag{7}$$

Inserting the right-hand sides of equations (7) and (3) into equation (2) yields the formula for Scott's π:

$$\pi = 1 - \frac{\sum_{i \neq j} p_{ij}}{\sum_{i \neq j} \left(\frac{p_{i\cdot} + p_{\cdot i}}{2}\right)\left(\frac{p_{j\cdot} + p_{\cdot j}}{2}\right)}. \tag{8}$$

Unlike *S* and Scott's π, Cohen's κ does not make the assumption of marginal homogeneity (Zwick 1988). Variation in states' propensity to form ties is not taken as a sign of dissimilarity but considered to be due to causes unrelated to the choice of tie partner. The calculation of κ's chance dissimilarity relies directly on the observed marginal proportions as best guesses for the "true" marginal proportions: $m_{i\cdot} = p_{i\cdot}$ and $m_{\cdot j} = p_{\cdot j}$. Inserting these terms into equation (4) yields the following formula for the chance correction of κ:

$$D_e^{\kappa} = \sum_{i \neq j} p_{i\cdot} p_{\cdot j}. \tag{9}$$

Inserting the right-hand side of equations (9) and (3) into equation (2) gives the formula for Cohen's κ:

$$\kappa = 1 - \frac{\sum_{i \neq j} p_{ij}}{\sum_{i \neq j} p_{i\cdot} p_{\cdot j}}. \tag{10}$$

The calculation of the different measures and the effects of prevalence and bias are best illustrated through a few examples. Fig. 4 illustrates the computation of the similarity measures in the absence of bias and prevalence. In this hypothetical example, alliance ties are just as common as non-alliance ties and both dyad members have the same propensity to form alliance ties. The observed proportion of dissimilarity is calculated by adding up the proportions in the off-diagonal cells of the contingency table:

[10] In the case of valued data, the standardization is equivalent to assuming that all states have a 50% propensity to form a tie with the theoretically possible maximum strength and a 50% propensity to form a tie with the theoretically possible minimum strength.

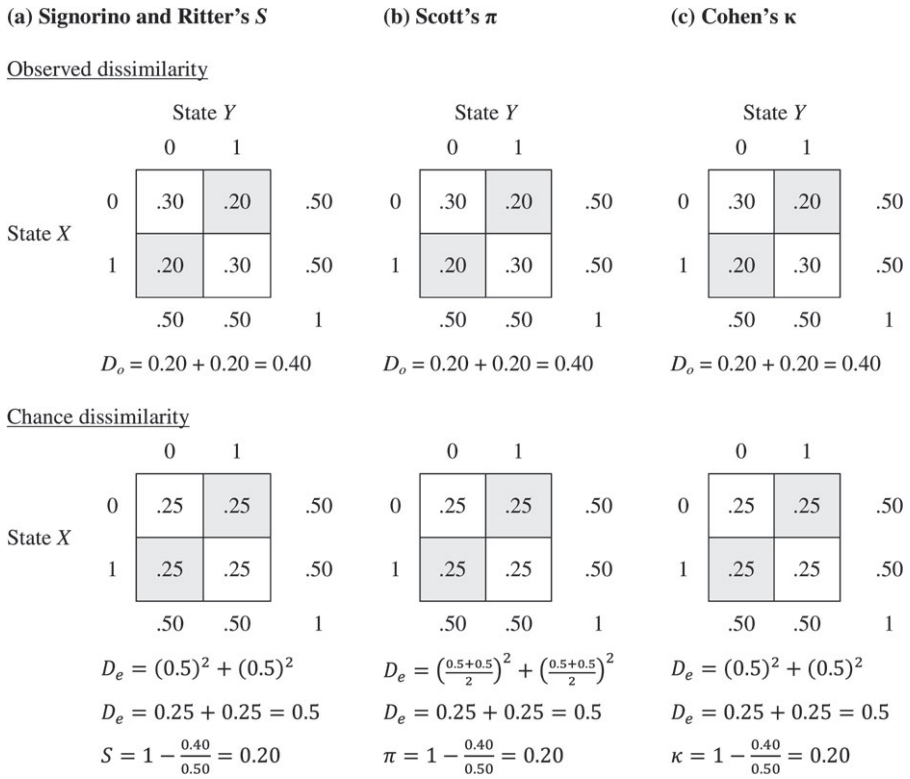

Fig. 4 The effect of chance correction in the absence of prevalence and bias.

$$D_o = \sum_{i \neq j} pij = p_{12} + p_{21} = 0.20 + 0.20 = 0.40.$$

The observed proportion of dissimilarity is always the same for all three similarity measures. However, in the absence of bias and prevalence, the chance dissimilarity is the same for all three measures as well. In this situation, the hypothesized marginal proportions are $m_{i\cdot} = m_{\cdot j} = 0.5$ regardless of the different assumptions about the ''true'' marginal distributions:

$$D_e^S = (0.5)^2 + (0.5)^2 = 0.25 + 0.25 = 0.5$$
$$D_e^\pi = \sum_{i \neq j} \left(\frac{p_{i\cdot} + p_{\cdot i}}{2}\right)\left(\frac{p_{j\cdot} + p_{\cdot j}}{2}\right) = \left(\frac{0.5+0.5}{2}\right)^2 + \left(\frac{0.5+0.5}{2}\right)^2 = 0.25 + 0.25 = 0.5$$
$$D_e^\kappa = \sum_{i \neq j} p_{i\cdot} p_{\cdot j} = (0.5)^2 + (0.5)^2 = 0.25 + 0.25 = 0.5.$$

As a consequence of the equality of the marginal proportions, the proportion of dissimilarity expected by chance is 0.5 for all three measures as well. In the example, the chance dissimilarity is slightly larger than the actually observed dissimilarity of 0.4, resulting in a moderately positive similarity score of 0.2.

If the marginal distributions are unbalanced in one way or another, the different indices give different results. An instance in which the data indicate the prevalence of non-alliance ties is illustrated in the upper half of Fig. 5. The proportion of dissimilarity is the same as in the previous example, but this time all shared ties are concentrated in the top left cell of the table indicating the absence of alliance commitments. The two states do not have a single alliance commitment in common. Given their marginal distributions, the two states' alliance portfolios are as dissimilar as they can possibly be. However, *S* does not take this information into account. The marginal proportions for the chance correction are calculated in exactly the

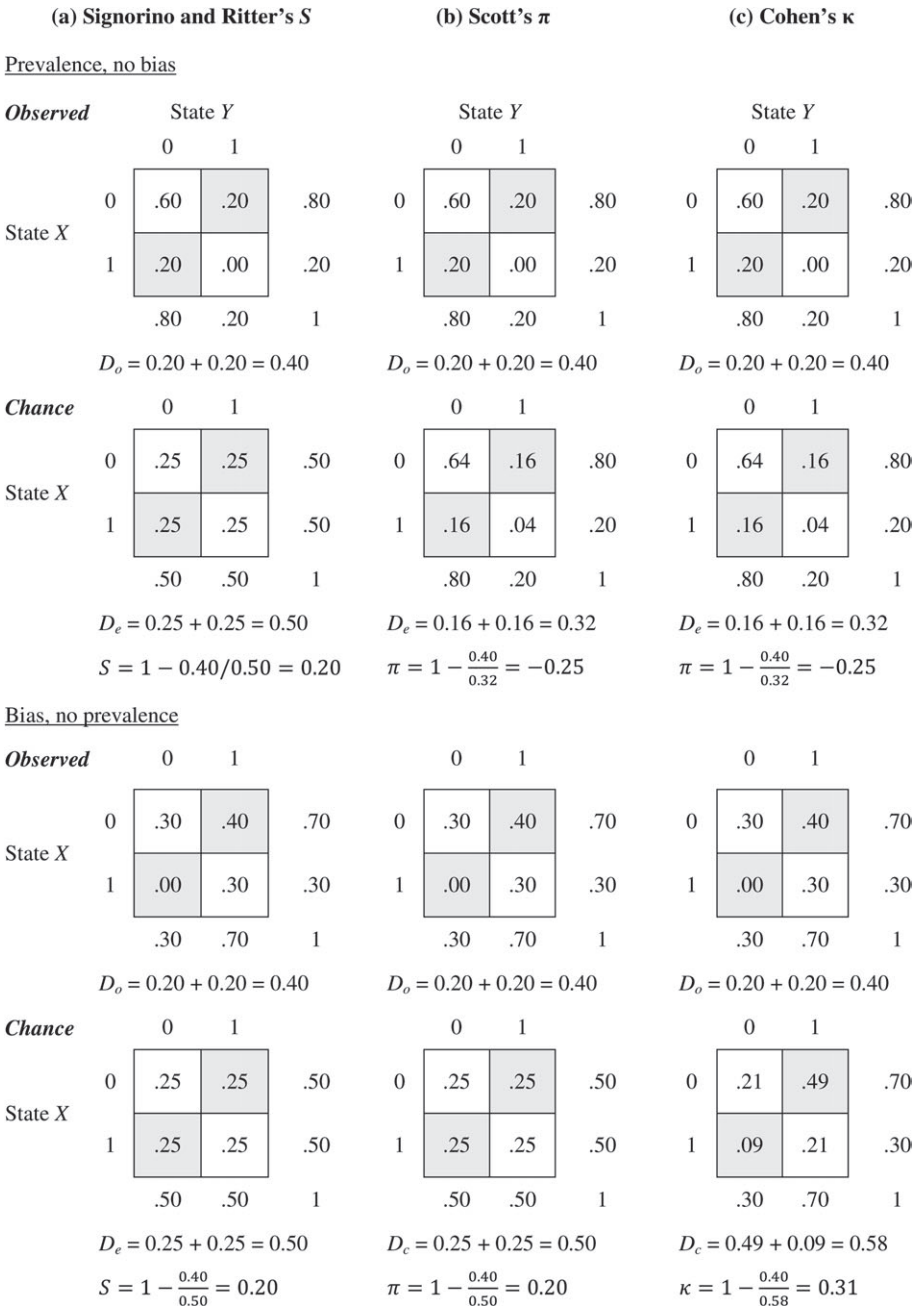

Fig. 5 The effect of chance correction in the presence of prevalence or bias.

same way as in the previous example, resulting in the same chance dissimilarity score of 0.5 and hence the same similarity score of 0.2. In contrast, the chance corrections of π and κ adjust their similarity scores for the fact that non-alliance ties are more frequent than alliance ties:

$$D_e^S = (0.5)^2 + (0.5)^2 = 0.25 + 0.25 = 0.5$$
$$D_e^\pi = \sum_{i \neq j} \left(\frac{p_{i\cdot} + p_{\cdot i}}{2}\right)\left(\frac{p_{j\cdot} + p_{\cdot j}}{2}\right) = \left(\frac{0.8+0.2}{2}\right)^2 + \left(\frac{0.2+0.8}{2}\right)^2 = 0.16 + 0.16 = 0.32$$
$$D_e^\kappa = \sum_{i \neq j} p_{i\cdot} p_{\cdot j} = (0.8)(0.2) + (0.2)(0.8) = 0.16 + 0.16 = 0.32.$$

Taking into account that a large dissimilarity score is more ''difficult'' to achieve when both dyad members have a large number of non-alliance ties, the proportion of dissimilar ties expected by chance reduces from 0.5 to 0.32 in the case of the agreement indices. This value is smaller than the observed proportion of dissimilarity of 0.4. As a consequence, the similarity scores of π and κ change from a moderate positive value of 0.2 to a moderate negative value of −0.25.

In the two examples considered so far, dyad members had identical marginal distributions, so π and κ yielded identical values. However, this observation changes when we consider the effect of bias. The lower half of Fig. 5 illustrates a situation in which the marginal distributions are unbalanced in a perfectly asymmetrical manner. State *X* has alliance commitments to only 30% of the other states in the system, but State *Y* has alliance commitments to 70% of the other states. Again, the proportion of dissimilarity is kept constant at 0.4, which means that *S* remains constant at 0.2. Yet by averaging the observed marginal proportions of dyad members to estimate the ''true'' marginal proportions, π also remains the same. Only κ takes bias into account by adjusting its similarity score for differential propensities of dyad members to form alliance commitments:

$$D_e^S = (0.5)^2+(0.5)^2 = 0.25+0.25 = 0.5$$
$$D_e^\pi = \sum_{i\neq j}\left(\tfrac{p_{i\cdot}+p_{\cdot i}}{2}\right)\left(\tfrac{p_{j\cdot}+p_{\cdot j}}{2}\right) = \left(\tfrac{0.7+0.3}{2}\right)^2+\left(\tfrac{0.3+0.7}{2}\right)^2 = 0.25+0.25 = 0.5$$
$$D_e^\kappa = \sum_{i\neq j} p_{i\cdot}p_{\cdot j} = (0.7)(0.7)+(0.3)(0.3) = 0.49+0.09 = 0.58.$$

When alliance ties of dyad members are asymmetrically distributed, then it is ''easier'' to exhibit a large proportion of dissimilar ties simply by chance. Only κ's chance correction takes this consideration into account, resulting in a higher chance dissimilarity of 0.58 compared to the chance dissimilarity of 0.5 of *S* and π. Correspondingly, κ's similarity value also increases from 0.2 to 0.31.

In summary, the chance correction model implicit in *S* expects that states will agree on 50% of their alliance ties by chance, regardless of the actually observed marginal distributions of the alliance commitments of the two dyad members. The *S* score is a simple linear transformation of the proportion of dissimilarity. In contrast, the calculation of π and κ does not rely solely on the observed proportion of dissimilarity. Their chance correction models also take information about the prevalence of certain types of ties into account when calculating similarity scores. For a given observed proportion of dissimilarity, their scores are lower the more prevalent a certain type of tie. In addition, κ also takes into account differential propensities of states to form alliance ties. For a given observed proportion of dissimilarity, the similarity score of κ is higher the larger the differences are between dyad members' marginal proportions. The value of Cohen's κ can be lower or higher than the value of *S*, depending on whether the effect of prevalence or the effect of bias outweighs the other (Lantz and Nebenzahl 1996, 434). As Scott's π adjusts *S* only downward, its similarity score is always the same or lower than the scores of *S* and κ.

4 Scott's π and Cohen's κ for Quantitative Data

Signorino and Ritter's *S* is routinely calculated on data with valued alliance ties.[11] The two chance-corrected agreement indices are readily extended to the case of interval-level data as well. Krippendorff (1970) provides formulations of π and κ in terms of two variables *X* and *Y*, representing the two tie profiles of the dyad members (see also Fay 2005). To measure the degree of dissimilarity between the two profiles, squared or absolute distances between tie values are often calculated (e.g., Shankar and Bangdiwala 2008, 447). However, the squared distances are usually preferred ''because of historical precedent, simplifications, and some nice properties'' (Fay 2005, 175; see also Krippendorff 1970, 141). Unless stated

[11] Although different types of alliance commitments are regularly treated as if they were ordered on an interval scale, this assumption is extremely questionable and I do not recommend relying on it. Unfortunately, most previous applications of *S* have treated the strength of alliance commitments as if they were based on quantitative data. Hence, for purely comparative reasons, I follow this practice in the remainder of this article.

otherwise, I follow this convention in the remainder of this article. The formulas for Scott's π and Cohen's κ for quantitative data take the following form:[12]

$$\pi = 1 - \frac{\sum (X-Y)^2}{\sum \left(X - \frac{\bar{X}+\bar{Y}}{2}\right)^2 + \sum \left(Y - \frac{\bar{X}+\bar{Y}}{2}\right)^2}. \tag{11}$$

$$\kappa = 1 - \frac{\sum (X-Y)^2}{\sum (X-\bar{X})^2 + \sum (Y-\bar{Y})^2 + \sum (\bar{X}-\bar{Y})^2}. \tag{12}$$

For comparative purposes, S can be expressed in a similar form:

$$S = 1 - \frac{\sum (X-Y)^2}{\frac{1}{2}\sum (d_{\max})^2} = 1 - \frac{\sum (X-Y)^2}{\sum \left(d_{\max} - \frac{d_{\max}}{2}\right)^2 + \sum \left(d_{\max} - \frac{d_{\max}}{2}\right)^2}. \tag{13}$$

In all three equations, the sum of squared distances in the numerator captures the dissimilarity in the scale values of foreign policy ties. Like in the case of binary ties, the three formulas differ only in the calculation of the dissimilarity expected by chance, which is given in the denominator. Again, the denominator of S is equivalent to the expectation that half of the theoretically possible maximum dissimilarity will occur by chance. The last expression in equation (20) demonstrates that this chance dissimilarity is equivalent to the sum of the theoretically possible maximum variability of each dyad member's valued tie profile.

The denominator of Scott's π consists of the sum of the observed variability of dyad members' valued tie profiles around the grand mean. The grand mean is simply the average of the two profile-specific means. Calculating the deviations from the grand mean rather than the profile-specific means reflects the assumption of homogenous marginal distributions. Unlike the chance correction of S, which uses the mid-point of the scale as the ''grand mean,'' the chance correction of π uses an empirical estimate of the grand mean. In this way, π takes into account that the two distributions might be symmetrically unbalanced or skewed in a similar way toward one or the other end of the scale. In contrast, the denominator of Cohen's κ assumes that chance dissimilarity is equal to the sum of the variability in the two dyad members' valued tie profiles plus the difference in their means. The variability of each tie profile is calculated around its profile-specific mean, implying that no assumption is made that the dyad members' propensity to establish foreign policy ties is identical. Adding the sum of the squared distances of the profile-specific means to κ's denominator indicates that asymmetrically unbalanced distributions are not considered to be a source of dissimilarity. On the contrary, the larger denominator directly results in a larger similarity score. To summarize, when ties are valued, prevalence takes the form of both dyad members having mean tie strength values similarly larger or lower than the mid-point of the scale, and bias takes the form of dyad members differing in their mean tie strength values.

5 The Empirical Consequences of Chance Correction

Up to this point, I have discussed the conceptual differences between Signorino and Ritter's (1999) S and the two chance-corrected agreement indices. For applications of those measures, it is important to assess whether the use of chance-corrected agreement indices is likely to lead to different empirical similarity values and different results of statistical analyses. For this purpose, I first present the distribution of values of the different similarity measures before turning to a replication of Gartzke's (2007) study of the ''Capitalist Peace.''

[12] Valued alliance data of a dyad can be represented in vector form, similar to the binary data example depicted in Fig. 2a. The only difference is that the vector entries are not restricted to 0s and 1s, but can range from 0 = ''no commitment,'' 1 = ''entente,'' and 2 = ''neutrality or nonaggression pact'' to 3 = ''defense pact.'' All elements of equations (11) to (13) can be directly calculated from the information in those vectors.

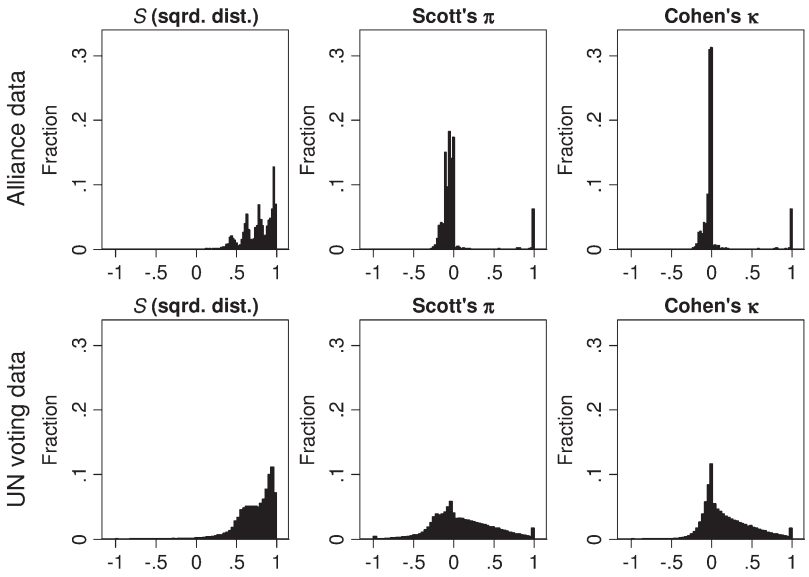

Fig. 6 Empirical distributions of similarity measures. The similarity measures in the first row of panels are based on alliance data of the Correlates of War Project (2003). The measures in the second row of panels are based on data of voting in the U.N. General Assembly (Voeten and Merdzanovic 2009).

Figure 6 compares the empirical distributions of S, π, and κ. The first row of panels is based on valued alliance data (Correlates of War Project 2003) and the second row on valued voting data from the U.N. General Assembly (Voeten and Merdzanovic 2009).[13] In both rows, the left panel shows that S scores clearly tend toward the upper end of the similarity scale (see also Bennett and Rupert 2003, 374). Taking these values literally, they indicate that almost all dyads have more common than diverging foreign policy positions. Given the observed preponderance of conflict in the international system, this seems to be a rather unrealistic description. In contrast, the distributions of both π in the middle column and κ in the right column are centered on zero.

To investigate the consequences of replacing S by π or κ in statistical analyses, I replicate Gartzke's (2007) study of the determinants of international conflict in the post-World War II era. Gartzke's (2007, 166) main argument is that "economic development, capital market integration, and the compatibility of foreign policy preferences," rather than joint democracy, account for the "dyadic democratic peace." Relying on logistic regression with a specification similar to Oneal and Russett's (1999) as a baseline model, Gartzke's reported results consistently show that the effects of democracy variables reduce in size and become statistically insignificant after adding liberal economic variables to the model.

Gartzke uses three different conflict measures as dependent variables: militarized interstate disputes, wars, and fatal militarized interstate disputes. The following discussion focuses on his analysis of the onset of war (Model 7 in Gartzke's Table 2) as the replication of this analysis resulted in the most disparate findings. The analysis of war onset is also of particular substantive importance, "as in some respects the most robust formulation of the democratic peace involves war" (Gartzke 2007, 179). Despite the prominence of the role of similar interests in Gartzke's theoretical argument, he includes the foreign policy similarity variable only in the analysis of all militarized interstate disputes, not in the analysis of fatal militarized disputes or the analysis of wars. Gartzke (2007, 180) explains that the variable is omitted "because it is not statistically significant in these regressions." Although my replication results for fatal militarized disputes (Model 9 in Gartzke's Table 2) are consistent with this claim, my replications results for war cannot reproduce this finding. In fact, the replication results show a statistically significant positive

[13]The U.N. voting variable distinguishes three values: 1 = "Yes," 2 = "Abstain," and 3 = "No." Based on these data, having voted in a similar way on the same U.N. resolutions determines the degree to which dyad members' foreign policy positions are judged to be similar.

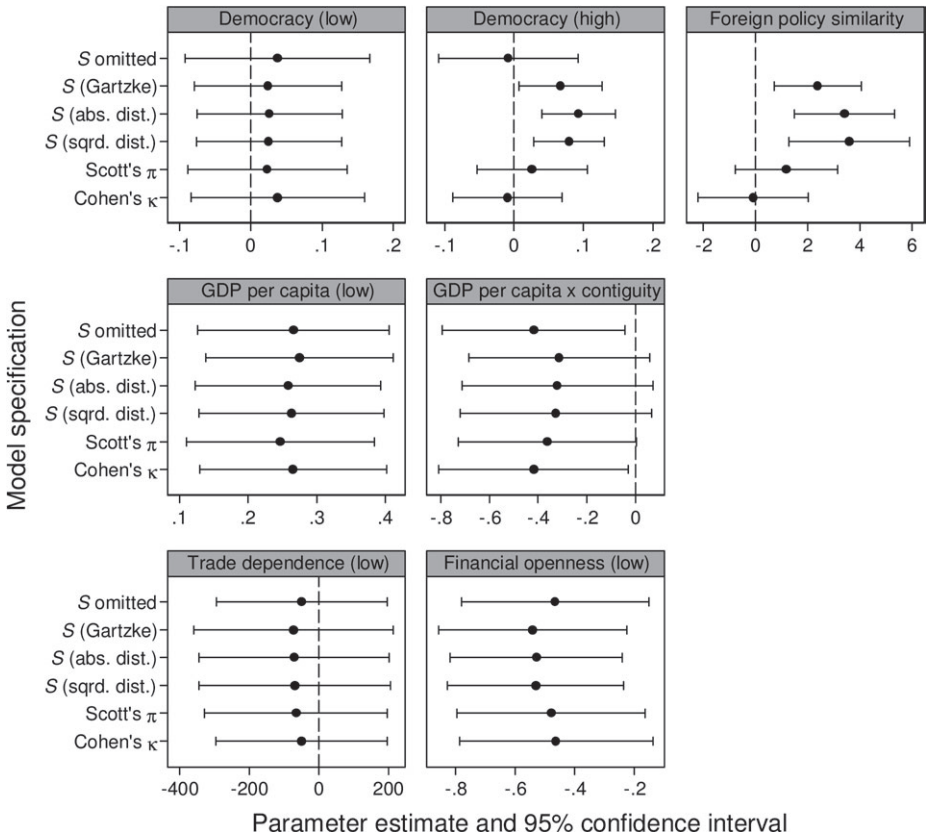

Fig. 7 The determinants of war: replication of Gartzke's (2007) "Capitalist Peace." The figure provides logistic regression coefficients and 95% confidence intervals for the theoretically interesting variables of Model 7 in Table 2 of Gartzke (2007, 181). The dependent variable is the onset of war. The first model specification excludes any measure of similarity and is a direct replication of Gartzke's results. The second model specification includes Gartzke's original measure of *S*, which is based on two-valued U.N. General Assembly voting data that treats abstentions as missing values. The remaining model specifications employ similarity measures that are based on three-valued U.N. voting data. The third specification includes *S* based on absolute distances and the fourth includes *S* based on squared distances. The last two specifications include Scott's π and Cohen's κ, both based on squared distances. The complete numerical regression results of the replication study are provided in Table A1 in the Appendix.

rather than the expected negative effect of foreign policy similarity on the probability of war onset. Even more problematically, the statistical inferences about two other explanatory variables change once Gartzke's *S* measure is included in the analysis.

Figure 7 illustrates the changes in regression coefficients and corresponding 95% confidence intervals resulting from changes in the way foreign policy similarity is or is not included in Gartzke's Model 7. The first model specification (*S* omitted) replicates Gartzke's original regression results excluding any measure of foreign policy similarity. The second model specification (*S* [Gartzke]) introduces Gartzke's own *S* measure, which is based on two-valued U.N. voting data, treating abstentions as missing values. The online appendix available from the *Political Analysis* web site reports replication results of two further studies that used *S* measures based on alliance data. Treating abstentions as missing values results in a loss of a lot of information as the abstention by one of the two dyad members is sufficient to discard the entire vote from the computation of the similarity measure. Thus, the computations for my similarity measures are based on three-valued voting data. The third and fourth models both include *S* measures but based on different distance metrics. This comparison allows us to distinguish between the effect of the distance metric and more qualitative properties of the similarity measures. The third model specification (*S* [abs. dist.]) includes *S* based on absolute distances and the fourth (*S* [sqrd. dist.]) includes *S* based on squared distances. Finally, the last two model specifications employ Scott's τ and Cohen's κ, both based on squared distance metrics.

Each panel in Fig. 7 presents the replication results for one of the main explanatory variables in the regression model. The absence of statistically significant effects of the democracy variables and the trade dependence variable, a statistically significant negative effect of financial openness, a positive effect of GDP per capita, and a negative effect of the interaction between GDP per capita and contiguity emerge as a consistent finding from Gartzke's study. However, this general finding does not hold for the regression analysis of war onset once *S* is included in the model specification. The regression coefficient for *S* is statistically significant, but its sign points in the "wrong" direction. More seriously, the inclusion of *S* changes the statistical inferences about two of the other six explanatory variables. The interaction term of GDP per capita with contiguity loses its statistical significance, whereas the variable recording the higher of the two democracy scores of the dyad members becomes statistically significant. The statistically positive effect of the latter variable contradicts Gartzke's core argument that regime type variables lose their explanatory power after controlling for liberal economic variables. As Fig. 7 shows, only model specifications that replace *S* by τ or κ can reproduce and bolster Gartzke's original claim. Thus, the replication results demonstrate that the choice of similarity measure can have profound impacts on the conclusions drawn from a statistical analysis.

6 The Proof of the Pudding Is in the Eating

Chance-corrected agreement indices like Scott's π and Cohen's κ have several desirable properties for measuring the similarity of state's foreign policy positions in the dyadic analysis of international relations. Although they assess the dissimilarity of dyad members' foreign policy tie profiles in a similar manner as Signorino and Ritter's (1999) *S*, π and κ differ crucially to *S* in the way they standardize the degree of dissimilarity. *S* relies on a standardization method that is equivalent to a rather arbitrary chance correction method, which will usually yield implausibly high similarity values. In contrast, the chance corrections of π and κ are based on the actually observed, empirical distributions of dyad members' foreign policy ties. Both measures adjust the similarity score for states' generally low propensity to form foreign policy ties. In addition, κ also adjusts the similarity score for differences in the individual propensities of states to form foreign policy ties. Whether or not the latter correction is reasonable depends mainly on the process supposed to generate the foreign policy tie data. If the data consist of alliance commitments, then the costs of ties are rather large and the assumption that states have different propensities to establish such ties seems reasonable. In this case, Cohen's κ is more appropriate than Scott's π. If the data consist of ties that are cheap to establish, such as a similar vote in the General Assembly of the U.N., then the assumption that all states have the same propensity to form a tie might be justified. In such a situation, differences in tie formation reflect real differences in foreign policy positions and π is preferable to κ as a measure of similarity.

An empirical comparison of similarity values shows clear differences between *S* on the one hand and the two chance-corrected agreement indices on the other hand. Although *S* scores tend toward 1, π and κ scores are concentrated around zero. In addition, a replication analysis of Gartzke's (2007) study of the "Capitalist Peace" demonstrates that the replacement of *S* by Scott's π or Cohen's κ can lead to different statistical inferences, not only about the effect of foreign policy similarity itself but also about the effects of other explanatory variables estimated in the same regression model.

Representing the similarity of two vectors in a single number is a surprisingly complex problem. Every similarity measure has its strengths and weaknesses and none is in any meaningful way wrong. Similarity measures represent exactly what they are mathematically defined to represent and good arguments can usually be found to advocate the use of each of them. In the end, the main yardstick of any similarity measure is the plausibility of the scores it produces. It is in this sense that "the proof of the pudding is in the eating." Although face validity is an especially shaky standard, most informed observers will agree that the U.K.'s foreign policy positions during the Cold War were considerably more similar to the positions of the United States than the positions of the Soviet Union. Although the scores for Scott's π and Cohen's κ reflect this consensus, *S* produces similarity values for the U.K.-Soviet Union dyad that are just below or above the similarity values of the U.K.-U.S. dyad (Fig. 1). The preceding discussion has shown that these implausible *S* scores are not due to exceptional circumstances that only affect this example, but result from general features built into the standardization of dissimilarity scores in *S*. The generality of the problem is also illustrated by the distribution of *S* scores (Fig. 6). Taking these similarity values literally, states hardly ever have serious differences in opinion, a description that seems to be contradicted by much of human history since the invention of the modern state.

https://doi.org/10.1093/pan/mpr023 Published online by Cambridge University Press

Appendix

Table A1 Replication of Gartzke's (2007, 181) logistic regression analysis of war onset (Model 7 in Table 2) using different measures of foreign policy similarity

	S omitted	*S Gartzke*	*S absolute distance*	*S squared dist.*	*Scott's τ*	*Cohen's κ*
Foreign policy similarity		2.386* (0.847)	3.412* (0.986)	3.595* (1.177)	1.193 (1.003)	−0.078 (1.076)
Democracy (low)	0.037 (0.066)	0.023 (0.053)	0.026 (0.052)	0.025 (0.052)	0.023 (0.057)	0.037 (0.062)
Democracy (high)	−0.008 (0.051)	0.067* (0.031)	0.093* (0.027)	0.080* (0.026)	0.026 (0.040)	−0.009 (0.040)
Trade dependence (low)	−48.757 (125.102)	−72.101 (146.172)	−70.511 (139.441)	−69.576 (140.367)	−65.884 (133.993)	−49.220 (125.668)
Financial openness (low)	−0.464* (0.161)	−0.540* (0.161)	−0.528* (0.148)	−0.530* (0.151)	−0.479* (0.161)	−0.462* (0.165)
GDP per capita (low)	0.266* (0.071)	0.275* (0.070)	0.258* (0.069)	0.263* (0.069)	0.247* (0.070)	0.266* (0.070)
GDP per capita × contiguity	−0.419* (0.192)	−0.312 (0.189)	−0.321 (0.199)	−0.328 (0.200)	−0.362 (0.187)	−0.418* (0.198)
Contiguity	4.655* (0.719)	4.382* (0.828)	4.370* (0.845)	4.397* (0.846)	4.527* (0.760)	4.630* (0.718)
Distance	−0.294 (0.191)	−0.516* (0.210)	−0.549* (0.213)	−0.532* (0.217)	−0.370 (0.221)	−0.293 (0.201)
Major power	1.550 (1.257)	2.678* (1.209)	2.905* (1.258)	2.697* (1.232)	1.871 (1.303)	1.571 (1.292)
Alliance	−1.053 (0.644)	−1.246* (0.530)	−1.248* (0.514)	−1.255* (0.523)	−1.305 (0.709)	−1.052 (0.771)
Capability ratio	−0.695* (0.229)	−0.895* (0.214)	−0.913* (0.212)	−0.900* (0.217)	−0.749* (0.215)	−0.691* (0.229)
North America	1.087 (0.951)	0.396 (0.961)	0.371 (0.992)	0.379 (0.984)	0.734 (0.902)	1.102 (0.932)
Africa	0.629 (0.952)	0.765 (0.938)	0.763 (0.921)	0.757 (0.929)	0.725 (0.931)	0.646 (0.949)
Middle East	2.364* (0.634)	2.522* (0.700)	2.507* (0.708)	2.530* (0.694)	2.369* (0.652)	2.363* (0.636)
Asia	−0.234 (0.843)	−0.747 (0.790)	−1.063 (0.764)	−0.904 (0.760)	−0.493 (0.819)	−0.114 (0.847)
Number of dyads	165,194	159,998	158,022	158,022	158,022	158,022
Log likelihood	−180.73	−173.99	−171.17	−172.93	−178.61	−180.43
Chi-square	312.06*	418.76*	392.16*	429.83*	338.92*	334.03*

Note. The table reports logistic regression coefficients with standard errors in parentheses. Estimation results for the constant term and the temporal splines are omitted. Regional dummies for Europe and South America were automatically dropped from the analysis.
*$p < 0.05$ (two-tailed significance tests).

In this respect, Scott's π and Cohen's κ offer very attractive alternative options for measuring the similarity of foreign policy positions and should be a valuable addition to the researcher's toolkit.

References

Altfeld, Michael F., and Bruce Bueno de Mesquita. 1979. Choosing sides in wars. *International Studies Quarterly* 23:87–112.

Bapat, Navin A. 2007. The internationalization of terrorist campaigns. *Conflict Management and Peace Science* 24:265–80.

Bearce, David H., Kristen M. Flanagan, and Katharine M. Floros. 2006. Alliances, internal information, and military conflict among member-states. *International Organization* 60:595–625.

Bennett, D. Scott, and Matthew C. Rupert. 2003. Comparing measures of political similarity: An empirical comparison of *S* versus in the study of international conflict. *Journal of Conflict Resolution* 47:367–93.

Braumoeller, Bear F. 2008. Systemic politics and the origins of Great Power conflict. *American Political Science Review* 102:77–93.

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

Byrt, Ted, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46:423–9.

Cicchetti, Domenic V., and Alvan R. Feinstein. 1990. High agreement but low kappa II: Resolving the paradoxes. *Journal of Clinical Epidemiology* 43:551–8.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46.

———. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70:213–20.

Correlates of War Project. 2003. *Formal interstate alliance dataset, 1816–2000. Version 3.03*. http://www.correlatesofwar.org/COW2%20Data/Alliances/Alliance_v3.03_dyadic.zip (accessed June 12, 2009).

———. 2005. *National material capabilities dataset." Version 3.02*. http://www.correlatesofwar.org/COW2%20Data/Capabilities/NMC_3.02.csv (accessed June 12, 2009).

———. 2005. *State system membership list. Version 2004.1*. http://correlatesofwar.org/COW2%20Data/SystemMembership/system2004.csv (accessed January 8, 2008).

De Vaus, David A. 2001. *Research design in social research*. London: Sage.

Derouen, Karl, and Uk Heo. 2004. Reward, punishment or inducement? US economic and military aid, 1946–1996. *Defence and Peace Economics* 15:453–70.

Fay, Michael P. 2005. Random marginal agreement coefficients: Rethinking the adjustment for chance when measuring agreement. *Biostatistics* 6:171–80.

Feinstein, Alvan R., and Domenic V. Cicchetti. 1990. High agreement but low Kappa I: The problems of two paradoxes. *Journal of Clinical Epidemiology* 43:543–9.

Gartzke, Erik. 1998. Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace. *American Journal of Political Science* 42:1–27.

———. 2007. The capitalist peace. *American Journal of Political Science* 51:166–91.

Kastner, Scott L. 2007. When do conflicting political relations affect international trade? *Journal of Conflict Resolution* 51:664–88.

Kellstedt, Paul M., and Guy D. Whitten. 2008. *The fundamentals of political science research*. Cambridge: Cambridge University Press.

Kendall, Maurice G. 1938. A new measure of rank correlation. *Biometrika* 30:81–93.

Kirk, Jennifer. 2010. *'rmac': Calculate RMAC or FMAC agreement coefficients." R package, Version 0.9*. http://cran.r-project.org/web/packages/rmac/ (accessed May 3, 2011).

Krippendorff, Klaus. 1970. Bivariate agreement coefficients for reliability of data. *Sociological Methodology* 2:139–50.

———. 2004. Measuring the reliability of qualitative text analysis data. *Quality and Quantity* 38:787–800.

Lantz, Charles A., and Elliott Nebenzahl. 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology* 49:431–4.

Lin, Lawrence I.-Kuei. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–68.

Long, Andrew G., and Brett Ashley Leeds. 2006. Trading for security: military alliances and economic agreements. *Journal of Peace Research* 43:433–51.

Morrow, James D., Randolph M. Siverson, and Tressa E. Tabares. 1998. The political determinants of international trade: the major powers, 1907–90. *American Political Science Review* 92:649.

Neumayer, Eric. 2003. What factors determine the allocation of aid by Arab countries and multilateral agencies? *Journal of Development Studies* 39:134–47.

Oneal, John R., and Bruce Russett. 1999. Assessing the liberal peace with alternative specifications: Trade still reduces conflict. *Journal of Peace Research* 36:423–42.

R Development Core Team. 2011. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Scott, John. 2000. *Social network analysis: A handbook*. London: Sage.

Scott, William A. 1955. Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly* 19(3):321–5.

Shankar, Viswanathan, and Shrikant I. Bangdiwala. 2008. Behavior of agreement measures in the presence of zero cells and biased marginal distributions. *Journal of Applied Statistics* 35:445–64.

Signorino, Curtis S., and Jeffrey M. Ritter. 1999. Tau-b or not tau-b: Measuring the similarity of foreign policy positions. *International Studies Quarterly* 43:115–44.

Sim, Julius, and Chris C. Wright. 2005. The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85:257–68.

Stone, R. W. 2004. The political economy of IMF lending in Africa. *American Political Science Review* 98:577–91.

Sweeney, Kevin, and Omar M. G. Keshk. 2005. The similarity of states: Using S to compute dyadic interest similarity. *Conflict Management and Peace Science* 22:165–87.

Vach, Werner. 2005. The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology* 58:655–61.

Voeten, Eric, and Adis Merdzanovic. 2009. *United Nations General Assembly voting data*. hdl:1902.1/12379UNF:3:Hpf6qOk-DdzzvXF9m66yLTg==.http://dvn.iq.harvard.edu/dvn/dv/Voeten/faces/study/StudyPage.xhtml?studyId=38311&;studyListing-Index=0_dee53f12c760141b21c251525331 (accessed June 12, 2009).

Zegers, Frits. 1986. A family of chance-corrected association coefficients for metric scales. *Psychometrika* 51:559–62.

Zwick, Rebecca. 1988. Another look at interrater agreement. *Psychological Bulletin* 103:374–8.