



# International Interactions

Empirical and Theoretical Research in International Relations



ISSN: 0305-0629 (Print) 1547-7444 (Online) Journal homepage: <https://www.tandfonline.com/loi/gini20>


## Consensus Decisions and Similarity Measures in International Organizations

Frank Häge & Simon Hug


To cite this article: Frank Häge & Simon Hug (2016) Consensus Decisions and Similarity Measures in International Organizations, *International Interactions*, 42:3, 503-529, DOI: [10.1080/03050629.2016.1138107](https://doi.org/10.1080/03050629.2016.1138107)

To link to this article: <https://doi.org/10.1080/03050629.2016.1138107>

 View supplementary material 



 Published online: 08 Apr 2016.

 Submit your article to this journal 

 Article views: 573

 View related articles 

 View Crossmark data 

 Citing articles: 6 View citing articles 

that report agreement over and beyond the agreement expected based on certain assumptions about the marginal vote distribution of dyad members.

Second, Bailey et al. (2013) convincingly show that currently used affinity measures cannot address the issue of changing agendas. More specifically, if due to a particular conflict a series of resolutions are voted upon in one year but not in the other, the preference configuration related to this conflict will strongly affect affinity measures even though the underlying preference similarity of states has not changed. According to these authors, a one-dimensional item-response theory (IRT) model with bridging observations across sessions formed by resolutions with very similar contents allows circumventing this problem.

A third issue, however, has so far remained largely unaddressed: the fact that consensus voting plays an important role in many international organizations in general and the UNGA in particular. In the UNGA, for instance, only a small share of resolutions are actually voted upon, while a large majority is adopted without a vote through a consensus decision.<sup>1</sup> Existing affinity measures and IRT-models rely exclusively on data about roll-call votes. Resolutions adopted without a vote are not reflected in these measures. As the share of resolutions adopted without a vote varies over time and also across issue domains (Hug 2012), both affinity measures and estimates from IRT models are affected by ignoring these missing “votes.”

In the present article, we discuss the issue of consensus decisions and show how it may be addressed in the context of studies using affinity scores.<sup>2</sup> We find that neglecting consensus decisions may seriously affect affinity values and inferences based on these measurements. More specifically, we replicate the study by Alesina and Dollar (2000) on the political and strategic factors explaining the allocation of bilateral aid by specific donors. We find that preference similarity as measured on the basis of UNGA votes fail to robustly affect aid allocation once we include information on consensus decisions and account for chance agreement.

In the next section we present a brief overview of research using affinity measures based on UNGA voting data. The section also highlights how the practice of consensus decision making might affect the results offered in these studies. In the following section we demonstrate in detail how chance agreement and consensus decisions (and their neglect) affect similarity measures. Next, we present a data set on UNGA voting, which for the first time comprises information about resolutions adopted without a vote. Through a replication of Alesina and Dollar’s (2000) study, the section shows that taking consensus decisions and the possibility of chance agreement into account is important for

---

<sup>1</sup>In this article, we will treat adoptions without a vote as synonymous with a consensus decision, as does much of the literature; see Blake and Lockwood Payton (2015).

<sup>2</sup>In the conclusion, based on some preliminary work, we offer some thoughts about how this problem might be addressed in the context of IRT models.

finding any effect of foreign policy preferences on aid allocation. Had Alesina and Dollar (2000) used Signorino and Ritter's (1999) popular *S* measure, while ignoring consensus votes, they would have concluded to no effect of these preferences. The last section concludes with a summary of the argument and study and some ideas for further research.

### Affinity measures and consensus decisions

Affinity measures have become very popular in quantitative analyses of various subfields in International Relations. For example, Gartzke (1998, 2007) draws heavily on them when dealing with explanations of interstate conflict. Alesina and Dollar (2000) have popularized these measures for the examination of strategic decisions of aid allocation. In terms of the exact measures employed, studies differ considerably. Alesina and Dollar (2000) rely simply on the proportion of common votes to identify to what degree a country is a friend of the United States (US) or Japan, while Gartzke (1998:14) employs Spearman's rho correlation coefficient. More recently Signorino and Ritter (1999) proposed a more sophisticated measure called *S*, which has subsequently become the standard for measuring state foreign policy preference similarity in International Relations research. Häge (2011) criticizes this measure because its scores are not adjusted for chance agreement that occurs for reasons other than preference similarity. As a solution, he proposes to use chance-corrected agreement indices instead.<sup>3</sup>

Bailey et al. (2013) propose another critique to these measures. They argue that over time the similarity measures are heavily influenced by agenda effects. If a particular conflict becomes important in a particular year, a series of votes will deal with it and thus emphasize a particular type of disagreement. This very same and persistent disagreement might not appear in the following year, simply because the conflict has subsided, and no resolutions address it anymore. Bailey et al. propose to overcome this problem by using an IRT model, which allows estimating ideal points based on observed voting decisions. In order to allow for changing preference configurations, the authors estimate ideal points for countries on a yearly basis but ensure that the scales of these ideal points are comparable by using very similar resolutions voted upon in several sessions as bridging observations from one session to the next. Consequently, changes in the location of ideal points can be considered as changes in preferences, and the distances among states give an indication of how close or far apart particular countries are from each other.<sup>4</sup> It is important to note that these bridging observations are only necessary if scholars wish to assess changes in similarity over time, as much of the literature does.

<sup>3</sup>See Stokman (1977) and Mokken and Stokman (1985) for similar suggestions in the context of UNGA voting.

<sup>4</sup>For a recent study using this measure, see Mattes, Leeds, and Carroll (2015).

However, this way to proceed is not without criticism, as the pertinence of the bridging observations is based on very strong assumptions. In particular, the approach assumes that the relevant policy space is one-dimensional and that the scale being estimated is the same from one year to the next. In addition, the ways in which ideal points translate into votes for the bridging observations are assumed to be the same over time as well. Jessee (2010) and Lewis and Tausanovitch (2013) assess some recent studies from the context of the American Congress employing a similar strategy. They find that the necessary assumptions are almost never fulfilled. In contrast, affinity measures can be derived as measures of similarity of foreign policy positions in a multidimensional space (Signorino and Ritter 1999). Importantly, the measures do not require the analyst to specify the number of dimensions in advance. Furthermore, the suggested chance-corrections moderate undesirable effects of changes in the agenda on dyadic similarity values. Thus, chance-correction addresses one of the major criticisms waged against simple dyadic similarity measures without making the arguably implausible assumption that a single and temporally stable dimension of contestation structures the international system.<sup>5</sup>

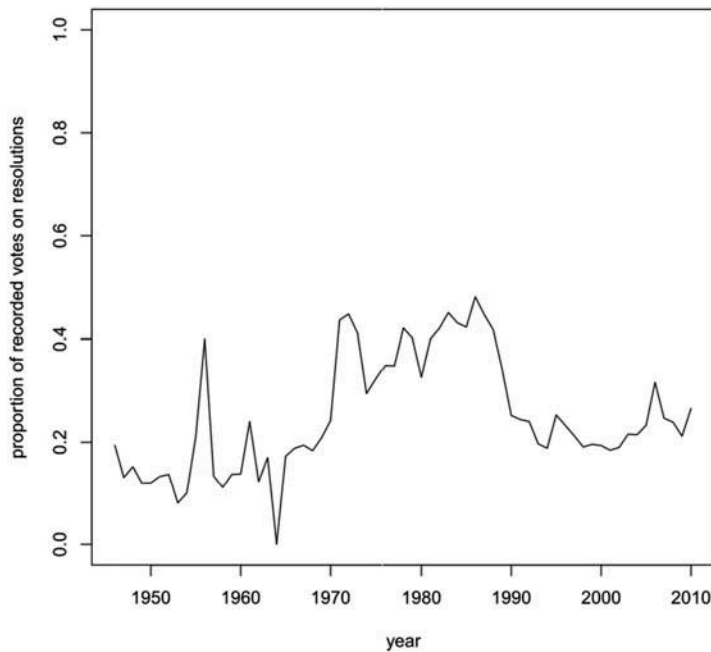
However, all of those measurement strategies take as basic input the roll-call votes in an assembly, usually the UNGA. This approach is problematic, as a large and variable share of UNGA resolutions are adopted without a formal vote.<sup>6</sup> While the large share of resolutions being adopted without a formal vote is acknowledged in the broader literature on the UNGA, its variation over time has been largely ignored. For the purpose of measuring the similarity of voting patterns, the existence of this variation over time has important implications. If always the same share of decisions were reached through consensus decisions, omitting those “votes” would still understate the similarity of voting patterns but would not affect the comparability of affinity values over time. However, if the share of consensus decisions varies, affinity measures that do not take consensus decisions into account cannot reasonably be compared over time. Figure 1 depicts the share of roll-call votes on final passage of UNGA resolutions in the period between 1945 and 2011 (Hug 2012). The figure shows that the share of roll-call votes has varied between a low of approximately 10% (with the exception of 1964) and a high of almost 50%. This implies that focusing only on roll-call votes ignores between 50% and 90% of all decisions on UNGA resolutions.<sup>7</sup>

---

<sup>5</sup>In selecting one or the other approach to measuring foreign policy similarity, researchers should consider to what extent they find this assumption justified.

<sup>6</sup>For discussions on voting rules in international organizations in general and consensus decision making in particular, see Blake and Lockwood Payton (2015). Presumably, the rationale for not taking consensus votes into account is that they do not provide for variation in voting behavior, but existing work does not explicitly justify or even discuss their exclusion (for example, Gartzke 1998; Alesina and Dollar 2000). We contend that consensus votes provide information about states' agreement and, as outlined in further detail below, that disregarding them leads to biased measures.

<sup>7</sup>Hug (2012) shows that there is considerable variation in the share of decisions adopted without a formal vote even in UNGA decisions not related to resolutions.



**Figure 1.** Proportion of roll-call votes on UNGA resolutions over time.

The problem generated by consensus decisions is akin to selection effects in roll-call vote analyses in parliaments (Hug 2010). We normally have very little guidance about how members of parliament voted in nonrecorded votes. However, in the case of decision-making bodies of international organizations, the lack of an explicit vote signals consensus among the delegates (Blake and Lockwood Payton (2015).<sup>8</sup> Consensus decisions in the UNGA are normally preceded, according to the minutes, by the chairperson asking whether a vote on a particular resolution (or any other matter) is necessary. Peterson (2005:3), in his discussion of changes in UNGA practices notes that this body “... also speeded deliberations on particular items through a set of unwritten practices for circulating drafts, presenting amendments or rival proposals, and developing a single draft through informal consultations held outside the public meetings.” Thus, he argues that by informal practices outside the regular sessions of the UNGA, a consensus is forged, implying that consensus actions are most likely akin to unanimous votes in favor.

<sup>8</sup>For some contested votes up to 1988, the UNGA’s minutes only report the marginal vote distribution rather than a full roll call. We refer to those votes as “nonrecorded,” as does the United Nations, to distinguish them from consensus decisions and roll-call votes. For the replication analyses reported in the main text, we omitted resolutions adopted through nonrecorded votes. However, in the Web appendix, we report the results of replication analyses based on the averages of five imputed data sets (as suggested by King, Honaker, Joseph, and Scheve 2001). More specifically, based on the reported marginal vote distributions, we randomly assigned yes and no votes as well as abstentions to the participating countries for all resolutions adopted through nonrecorded votes. The similarity measures were then calculated from the imputed data sets. The results show that these imputations barely affect our substantive conclusions, largely because the number of such nonrecorded votes has declined dramatically during the time period we cover. The number of nonrecorded votes are as follows (years not listed after 1970 had no such votes): 64 (1970), 49 (1971), 40 (1972), 33 (1973), 36 (1974), 31 (1975), 7 (1976), 9 (1977), 1 (1978), 2 (1979), 17 (1980), 13 (1981), 2 (1982), 2 (1984), and 1 (1988).

A possible criticism of our approach is that consensus decisions might simply relate to less-important resolutions. One way to assess this claim is to consider a commonly used source to identify salient UNGA decisions. Since the 1980s, the US State Department is by law required to offer a report on “Voting Practices in the United Nations,” in which it highlights the most important decisions and how UNGA members voted compared to the US. In the 1980s, the State Department chose for each session of the UNGA 10 roll-call votes, most of which are final passage votes of resolutions, that it deemed to be “key.”<sup>9</sup> However, in 1988, the State Department designated for the first time three decisions reached without a vote as equally important. That these resolutions are not innocuous is illustrated by the topic covered in the first consensus decision designated by the US State Department as being important, namely resolution 43/20 entitled “The Situation in Afghanistan and Its Implication for International Peace and Security.” This important resolution addressed concerns by UNGA members about the deteriorating situation in this war-torn country (see also Thacker 1999:73). Since then, the reports by the State Department list both important votes and important “consensus actions.” In Figure 2 we depict the number and share of important votes and consensus actions from 1983 to 2012.

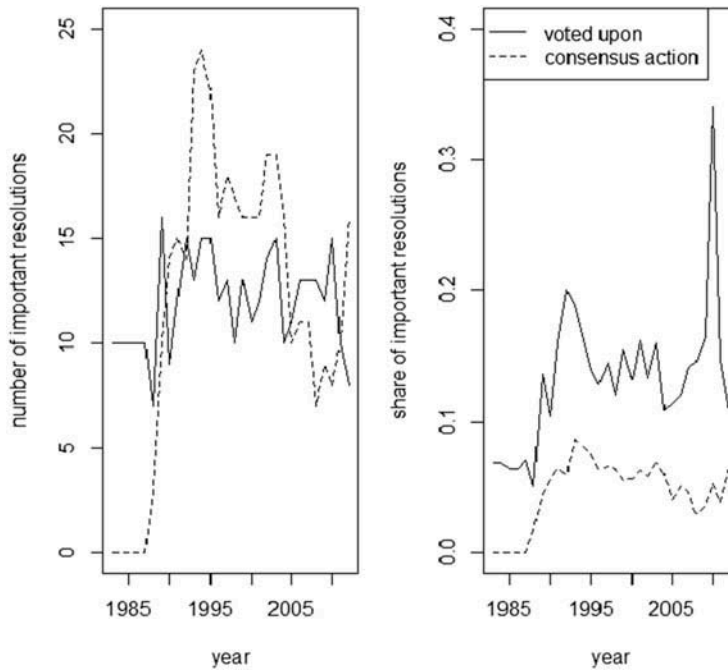
As the left panel of Figure 2 clearly shows, for large periods of time the US State Department considered more consensus decisions as important than matters adopted in a roll-call vote. In addition, the number of important votes and consensus actions do not evolve in parallel, suggesting again that variation across time is crucial and needs to be taken into account when assessing whether pairs of countries display similar preferences. When considering the share of roll-call votes and consensus decisions deemed important by the US State Department in the right panel of Figure 2, we note that the former share is always larger than the latter. However, the share of consensus votes is still of considerable size. In addition, the shares do not evolve in parallel over time. Thus, while consensus votes might be on average somewhat less important, the figure also shows that they are far from negligible.

### Accounting for consensus decisions in affinity measures

Having introduced the problem caused by consensus decisions and demonstrated their prevalence in the UNGA, we now turn to a more-detailed discussion about why consensus votes generate biases in affinity measures. For the purpose of this analysis, we take consensus votes at their face value and treat them as if all members of the UNGA explicitly voted *yea*.<sup>10</sup> Even if a formal vote was not

<sup>9</sup>For a list, see the appendix in Thacker (1999).

<sup>10</sup>The affinity measures are not affected by the way consensus votes are coded, as long as they are coded in the same way for all member states. Assuming that a consensus vote indicates either abstentions by all states or *no* votes by all states would lead to the same affinity score as assuming that it indicates *yes* votes by all states. However, the assumption that it signifies *yes* votes makes more substantive sense.



**Figure 2.** Number and share of important resolutions in the UNGA over time.

taken, consensus implies unanimous agreement, and member states are on the record for having supported the decision.<sup>11</sup> A number of reasons come to mind why the apparent support for a decision through a consensus vote might not be a true reflection of the actual position of a member state. For example, a member state might have budged under peer pressure or be responding to threats and promises of a more powerful state, the state might be engaged in a logroll of votes across resolutions, or the state might simply try to avoid being seen as having lost out in the negotiations for domestic reasons. However, in principle, all of these reasons for why consensus votes might not reflect the true position of a member state apply to more explicit *yea* votes as well.<sup>12</sup> If we treat explicit *yea* votes as being indicative of policy positions, little reason exists to treat implicit *yea* votes differently. In this respect, treating all consensus votes as *yea* votes is not an arbitrary auxiliary assumption but follows directly from the general logic of roll-call vote

<sup>11</sup>Indeed, in some international organizations where consensus voting is common, the respective decisions are explicitly recorded as having been adopted “by unanimity”. A prime example of this practice is the Council of the European Union (Häge 2013). In fact, an important reason for consensus decisions not being adopted through a roll call might be the actual absence of opposition to a motion. If it is clear from the outset that all states agree to a motion, taking a roll call is redundant.

<sup>12</sup>We acknowledge that, empirically, the incidence of extraneous factors being responsible for a *yea* vote might be higher in the case of consensus than recorded votes. However, the distinction between consensus and recorded votes in this respect is a matter of degree, not a qualitative one. An important rationale for applying a chance-correction is its adjustment of similarity scores for the possibility that covoting is not purely a result of similar policy positions. But again, although this correction might be somewhat more important when including consensus votes in the analysis, the same considerations apply equally when only roll-call votes are considered. Indeed, earlier proposals for applying chance-corrections to similarity indices were made in the context of analyses of roll-call votes only, see Mokken and Stokman (1985).



analysis that suggests that votes constitute revealed preferences of actors.<sup>13</sup> In the remainder of this section, we further elaborate on how the neglect of consensus votes in the calculation of vote agreement indices is justified neither on conceptual nor methodological grounds. We also illustrate how the neglect of consensus votes leads to generally biased agreement values as well as problems regarding their comparability over time.

### ***The effect of ignoring consensus votes on vote agreement measures***

A core component of most agreement measures is the proportion of disagreement. Of course, the proportion of disagreement is just the converse of the proportion of agreement. The latter is, for example, directly used to gauge interest similarity by Alesina and Dollar (2000).<sup>14</sup> However, the proportion of disagreement also lies at the heart of Ritter and Signorino's (1999) *S*, which is currently the standard measure used in the international relations literature to assess the similarity of states' UNGA voting profiles. In the case of a nominal variable, the proportion of disagreement is simply the sum of the proportion of observations falling in the off-diagonal cells of the contingency table of the UNGA voting variables of the two states. For  $i, j = 1, \dots, k$  nominal categories and  $p_{ij} = f_{ij}/f_{..}$  indicating the proportion of observations falling within cell  $ij$  of the contingency table, the proportion of disagreement is given by:

$$D_o = \sum_{i=1}^k \sum_{j=1}^k p_{ij} \text{ for } i \neq j \quad (1)$$

In the case of ordinal variables, the observations in the off-diagonal cells of the contingency table can be weighted to reflect varying degrees of disagreement (Cohen 1968). In the case of UNGA voting records, the voting behavior variable of each state can take three values: *yea*, *abstain*, and *nay*. Although these values reflect categories, most scholars assume them to be ordered along the dimension of support for the resolution voted upon (for example, Lijphart 1963:910; Gartzke 1998:14–15, but see Voeten 2000:193). Thus, weighting the

<sup>13</sup>One of the anonymous reviewers suggested that our proposal replaces the empirically untestable assumption that roll-call votes are representative of consensus votes with the equally untestable assumption that all states voted in favor when a resolution was adopted through a consensus vote. In our view, we are merely extending an already existing assumption made in analyses of recorded votes to consensus votes. In any case, our approach provides at least an alternative way of measuring preference similarity that broadens the methodological choice set for researchers. Where no state objected to the adoption of a resolution and is on public record for not doing so, it seems more plausible to us to assume that everybody was in favor of the resolution than to assume that 20%, 30%, or maybe even 40% of the states privately opposed the resolution but did prefer to not voice their dissent publicly (which is implied by the assumption that roll-call votes are representative of consensus votes). In general, consensus decision making seems to follow a similar logic in all international organizations. Thus, the choice between the two assumptions needs to be made on conceptual rather than empirical grounds.

<sup>14</sup>Agreement measures can either be formulated in terms of the proportion of agreement  $p^A$  or the proportion of disagreement  $p^D$ , where  $p^A = 1 - p^D$ . The choice of formulation is arbitrary. We focus on the proportion of disagreement, as it is equivalent to the "sum of distances" measures used to measure agreement in the case of interval-level variables.



**Table 1.** Calculation of Proportion of Disagreement for Ordinal Variables.

		State B		
		1 ( <i>Nay</i> )	2 ( <i>Abstain</i> )	3 ( <i>Yea</i> )
State A	1 ( <i>Nay</i> )	$p_{11}$ $w_{11} = 0$	$p_{12}$ $w_{12} = 1$	$p_{13}$ $w_{13} = 2$
	2 ( <i>Abstain</i> )	$p_{21}$ $w_{21} = 1$	$p_{22}$ $w_{22} = 0$	$p_{23}$ $w_{23} = 1$
	3 ( <i>Yea</i> )	$p_{31}$ $w_{31} = 2$	$p_{32}$ $w_{32} = 1$	$p_{33}$ $w_{33} = 0$
		$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$
				1

difference between a *yes* and a *no* vote heavier in the calculation of the proportion of disagreement than the difference between one of the extreme categories (that is, *yea* or *nay*) and the middle category (that is, *abstention*) seems justified. Table 1 illustrates this approach with a particular weighting function that assigns weights  $w_{ij}$  to cells according to the absolute difference between the row and column index number, that is,  $W_{ij} = |i - j|$ . This weighting is equivalent to treating the voting variables as exhibiting interval-level scales and calculating the absolute distance between the dyad members' variable values. The latter approach is taken in the calculation of disagreement values for  $S$ . We prefer the formulation in terms of disagreement weights, as it highlights that the precise degree to which different categories indicate disagreement is not given “naturally” by the values used to code those categories but needs to be the subject of a conscious decision by the researcher.<sup>15</sup> Taking weights for different degrees of disagreement into account and normalizing the sum of the weighted proportions by the maximum weight  $w_{\max}$ , the proportion of disagreement for ordered categories is given by the following formula:

$$D_o = \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{ij}}{w_{\max}} \quad (2)$$

The weights for the individual cells given our particular weighting function are shown in Table 1. For example, the weight for the “State A: nay, State B: abstain” cell ( $i = 1, j = 2$ ) is calculated by subtracting its column index number from its row index number and taking the absolute value of the resulting difference:  $w_{12} = |1 - 2| = |-1| = 1$ . The maximum weight is calculated by subtracting the highest row (column) index number from the smallest column (row) index number and taking the absolute difference. In our case, the index can take values from 1 to 3, hence  $w_{\max} = |3 - 1| = |1 - 3| = 2$ .

<sup>15</sup>For example, another prominent weighting function for ordered categorical data assigns weights to cells according to the squared distance between the row and column index number, that is,  $w_{ij} = (i - j)^2$ . Applying this weighting function is equivalent to calculating the squared distance between dyad members' variable values on interval-level scales. However, as no compelling reason exists to weight the difference between the two extreme categories four times heavier than the difference between the middle category and one of the extreme categories, we do not consider this weighting function in our analyses.

Table 2 shows how the UNGA voting information for the calculation of agreement values is usually represented in matrix format. Dyadic agreement values are calculated for each year based on the observed voting behavior of states on resolutions adopted during that time period.<sup>16</sup> The table presents data for 2 years, with 10 resolutions adopted in each of them, and information about the voting behavior of five major powers. While the table consists of artificial data constructed to illustrate our point about the detrimental effects of neglecting consensual decisions, the states and their values on the voting variables were chosen to roughly mirror the expected voting behavior of the five permanent UN Security Council members during the Cold War. During that period of time, the United States (USA) had diametrically opposed interests to the Soviet Union (USSR), the United Kingdom (UK) and France were more closely aligned with the USA, and China had more interests in common with the USSR.<sup>17</sup> The rows of the table with a gray background indicate resolutions adopted by consensus. Existing measures of vote agreement ignore these types of resolutions.

The arbitrariness of the neglect of consensus votes is best illustrated by considering the voting variable values of the USA and the USSR in year 1. Recall that the proportion of disagreement captures the degree to which dyad members' voting decisions differ from each other. The calculation of the proportion of disagreement relies exclusively on information about the voting behavior of the two states that are members of the particular dyad. In our example, only the information provided in the columns for the USA and USSR of Table 2 are of relevance for calculating the dyadic, year-specific vote agreement value for these two countries (as highlighted by the heavy-bordered rectangle). As the voting behavior of third parties is irrelevant for the calculation of the proportion of disagreement, no compelling reason exists to exclude resolutions on which both the USA and the USSR voted in favor, just because all other states voted in favor as well. Consider the first four resolutions of year 1. In all four cases, both the USA and the USSR voted in favor of the resolution. Yet when consensual decisions are excluded from the data set, the voting behavior on the first two resolutions is discarded. From a measurement point of view, given how the proportion of disagreement is defined, the voting behavior on the first two resolutions provides exactly the same information for the calculation of the proportion of disagreement between the USA and the USSR than the third and fourth resolution.

Ignoring resolutions adopted by consensus has nontrivial consequences for the agreement scores. First, given the large number of consensual decisions during a certain year, the agreement scores are generally biased downwards. Second, and possibly more important, agreement scores differ over time simply as a result of

<sup>16</sup>UNGA sessions and years do not completely overlap. As the temporal scope of the units of analysis usually used in international relations research is the year or a multiple thereof, we calculate agreement scores for individual years rather than UNGA sessions. In the calculation of dyadic similarity scores, a particular resolution is only included if both states were present during the meeting in which the resolution was adopted.

<sup>17</sup>The extent to which the artificial data in Table 1 do indeed reflect the actual voting behavior of those states during the Cold War is incidental to the argument we make here.

**Table 2.** The Structure of UN General Assembly Voting Data.

Year	Resolution	USA	USSR	UK	France	China
1	1	3	3	3	3	3
1	2	3	3	3	3	3
1	3	3	3	3	2	1
1	4	3	3	2	2	1
1	5	3	1	3	3	1
1	6	3	1	3	3	1
1	7	3	2	3	2	2
1	8	2	1	2	3	1
1	9	2	2	3	3	2
1	10	1	3	2	2	2
2	1	3	3	3	3	3
2	2	3	3	3	3	3
2	3	3	3	3	3	3
2	4	3	3	3	3	3
2	5	3	1	3	3	1
2	6	3	1	3	3	1
2	7	3	2	3	2	2
2	8	2	1	2	3	1
2	9	2	2	3	3	2
2	10	1	3	2	2	2

*Note.* The table presents artificial data constructed by the authors to resemble an extract from the UN General Assembly voting data for the five permanent UN Security Council members during the Cold War. The table includes data for two years with 10 resolutions adopted in each of them. The numerical codes of the voting variables indicate 1 = *nay*, 2 = *abstain*, and 3 = *yea*. The rows with a gray background indicate resolutions that have been adopted by consensus. The thick-lined rectangle indicates the voting information for the USA-USSR dyad. The illustration in the text of the calculation of various agreement measures focuses on this dyad.

the proportion of consensual decisions changing from year to year. Thus, discerning whether changes in dyadic agreement scores over time are really due to changes in the underlying voting profiles of states rather than changes in the proportion of consensual decisions becomes impossible. Table 3 illustrates these problems with our example data from Table 2. Each contingency table demonstrates the calculation of the proportion of disagreement between the USA and the USSR. The left column of contingency tables is based on the voting behavior in year 1 and the right column of contingency tables on the voting behavior in year 2. The first row of contingency tables shows the situation where consensual decisions are included in the calculation of the proportion of dissimilarity, while the second row illustrates the situation where they are excluded from the sample. To identify the effect of ignoring consensual decisions, the voting profile of each dyad member was constructed to be exactly the same in both sessions. The two sessions only vary in the number of consensual decisions taken, that is, in the way third states voted. In year 1, two out of 10 decisions (that is, 20%) were taken by consensus. In contrast, in year 2, four out of 10 decisions (that is, 40%) were taken by consensus. As Figure 1 indicates, these are rather conservative numbers given the often much higher consensus rates and fluctuations over time found in the real world.

Given that the voting profiles of the two states do not change from one session to the other, we would expect the proportion of disagreement to be the same as

**Table 3.** Consequences of Excluding Consensual Decisions.

Year 1					Year 2						
USA					USA						
	1	2	3	Total		1	2	3	Total		
A. Consensual Decisions Included											
USSR	1	0	1	2	3	USSR	1	0	1	2	3
		(.00)	(.10)	(.20)	(.30)			(.00)	(.10)	(.20)	(.30)
		0	1	2				0	1	2	
	2	0	1	1	2		2	0	1	1	2
		(.00)	(.10)	(.10)	(.20)			(.00)	(.10)	(.10)	(.20)
		1	0	1				1	0	1	
	3	1	0	4	5		3	1	0	4	5
		(.10)	(.00)	(.40)	(.50)			(.10)	(0)	(.40)	(.50)
		2	1	0				2	1	0	
Total		1	2	7	10	Total		1	2	7	10
		(.10)	(.20)	(.70)	(1)			(.10)	(.20)	(.70)	(1)
$D_0^1 = \frac{0.80}{2} = 0.4$					$D_0^1 = \frac{0.80}{2} = 0.4$						
USA					USA						
	1	2	3	Total		1	2	3	Total		
B. Consensual Decisions Excluded											
USSR	1	0	1	2	3	USSR	1	0	1	2	3
		(.00)	(.125)	(.25)	(.375)			(.00)	(.17)	(.33)	(.50)
		0	1	2				0	1	2	
	2	0	1	1	2		2	0	1	1	2
		(.00)	(.125)	(.125)	(.25)			(.00)	(.17)	(.17)	(.33)
		1	0	1				1	0	1	
	3	1	0	2	3		3	1	0	0	1
		(.125)	(.00)	(.25)	(.375)			(.17)	(.00)	(.00)	(.17)
		2	1	0				2	1	0	
Total		1	2	5	8	Total		1	2	3	6
		(.125)	(.25)	(.625)	(1)			(.17)	(.33)	(.50)	(1)
$D_0^1 = \frac{1}{2} = 0.5$					$D_0^1 = \frac{1.33}{2} = 0.67$						

Note. The tables are based on the artificial data presented in Table 2. The rows and columns of each table indicate the absolute and relative number of different types of votes (1 = *nay*, 2 = *abstain*, 3 = *yea*). The first figure of each cell gives the absolute number, the second figure in parentheses gives the proportion, and the third number gives the disagreement weight. The overall proportion of disagreement in voting can then be computed as the weighted sum of proportions divided by the maximum weight. For example, the proportion of disagreement for year 1 when consensual decisions are included in the calculation is computed by multiplying the third number with the second number in each cell of the table and adding up the resulting products. The sum of products is then divided by the maximum disagreement weight of 2:

$$D_0^1 = (0 * 0 + 1 * 0.1 + 2 * 0.2 + 1 * 0 + 0 * 0.1 + 1 * 0.1 + 2 * 0.1 + 1 * 0 + 0 * 0.4) / 2 = (0.1 + 0.4 + 0.1 + 0.2) / 2 = 0.8 / 2 = 0.4$$

well. Indeed, when consensual decisions are taken into account, the contingency tables for the two sessions are identical, and so are the associated values for the proportion of disagreement. When consensual decisions are ignored, the situation looks very different. The overall number of resolutions in each session is obviously reduced. Even though only the frequency of observations in the “3, 3” cell changes, the proportions for all cells increase as a result of the reduced number of

resolutions. Given that only the off-diagonal cells indicating disagreement receive non-zero weights in the calculation of the proportion of disagreement, the proportion of disagreement is generally larger when consensus votes are ignored than when they are included. In other words, if consensual decisions are ignored, measures based on the proportion of disagreement, including Ritter and Signorino's  $S$ , systematically understate vote agreement.

In this particular example, the proportion of disagreement is 0.40 in both years when consensual decisions are included.<sup>18</sup> In contrast, the proportion of disagreement is 0.50 in year 1 and 0.67 in year 2 when consensual decisions are excluded. The generally higher proportions of disagreement when consensual decisions are ignored illustrate the bias generated by their exclusion. The difference in the proportion of disagreement between 0.50 in year 1 and 0.67 in year 2 also shows how the proportion of disagreement varies simply as a result of different consensus rates. The two sessions indicate different proportion of disagreement scores even though the voting profiles of the two states are exactly the same. This finding highlights the more severe problem resulting from the exclusion of consensual decisions: proportions of disagreement scores are generally not comparable across time, as the size of the measurement bias varies with the size of the consensus rate. The larger the consensus rate of a particular session, the more agreement scores are biased toward more disagreement.<sup>19</sup>

### ***Correcting vote agreement for chance***

In its raw form, the proportion of disagreement will generally be very low if consensual decisions are taken into account. When the proportion of disagreement is rescaled to indicate agreement, measures relying on this quantity will indicate very high agreement scores. From a measurement point of

<sup>18</sup>See the notes to Table 3 for a detailed example of how the proportion of disagreement is calculated from the information in the contingency tables.

<sup>19</sup>Note that the size of the bias is not constant across dyads within a year. For example, consider a proportion of disagreement of 0.5 resulting from contrary voting on four out of eight roll-call votes. Adding two consensus votes increases the denominator from 8 to 10, resulting in a proportion of disagreement of 0.4. Now consider a proportion of disagreement of 0.25 resulting from contrary voting on two out of eight roll-call votes. Adding two consensus votes in this situation results in a proportion of disagreement of 0.2. Thus, whereas the bias in the first situation is 0.1, it is 0.05 in the latter. The situation becomes even more complicated when chance-corrections are applied, as the resulting similarity values are generally nonlinear functions of the proportion of disagreement. The differential impact on dyads within the same year implies that the bias resulting from ignoring consensus votes cannot be avoided by including control variables, such as time dummies or a continuous variable for the number of consensus votes, in statistical analyses. If the number of consensus votes was the same for each dyad in each year, including the number of consensus votes in that year plus its interaction with the proportion of disagreement could in principle be a technical substitute for including consensus votes in the measure itself. However, in practice, the number of consensus votes is not constant for all dyad members in a certain year. A dyadic similarity score can only be calculated if both dyad members participated in the adoption of a particular resolution. Due to some states only being members during part of a year or simply not attending the General Assembly meeting in which a resolution has been adopted, this is not always the case. As a result, the number of consensus votes varies from dyad to dyad. In fact, absenteeism is quite common in the General Assembly; roughly 89% of all dyad similarity scores are based on a number of resolutions that is lower than the total number of resolutions adopted during a particular year because one or both dyad members did not attend a the meeting in which a particular resolution was adopted. Furthermore, this percentage varies widely over time from 21% in 1955 to 100% in 1985.

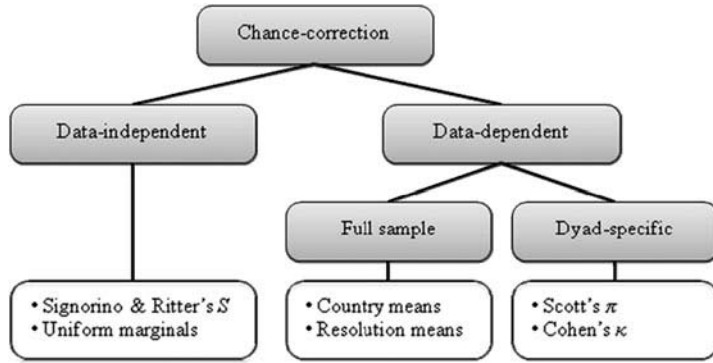
view, these high scores are not problematic, as they indicate exactly what the data tell us: most of the time, both dyad members support the adoption of a resolution. However, if we are interested in using vote agreement of states as an indicator for the similarity of their foreign policy preferences, we might want to compare the observed agreement to the agreement expected simply by chance. Of course, we are not suggesting that voting occurs randomly in real-life situations. However, the hypothetical scenario of random and independent voting provides a yardstick for making a judgment about the extent to which the actually observed agreement expresses similar or dissimilar policy positions (Stokman 1977:84). States with similar policy positions will have a much larger agreement score than the score expected by random voting, and states with dissimilar policy positions will have a much lower agreement score than the one expected by random voting. In addition, the agreement expected by chance is computed based on assumptions about the marginal distribution of probabilities with which states vote a certain way. Depending on the nature of the assumptions, they can alleviate concerns about agenda effects and covoting occurring due to factors other than preference similarity. In general, any chance-corrected agreement index  $A$  takes the following form:

$$A = 1 - \frac{D_o}{D_e} \quad (3)$$

The observed proportion of disagreement  $D_o$  is divided by the proportion of disagreement expected by chance  $D_e$ . The ratio is then subtracted from 1 to rescale the value to indicate the degree of agreement rather than disagreement. A value of 1 indicates perfect agreement, values between zero and 1 indicate more agreement than expected by chance, a value of zero indicates agreement no different from chance, and values below zero indicate more disagreement than expected by chance.

While the general structure of chance-corrected agreement indices is the same for all, they differ in their assumptions about the disagreement expected by chance. Broadly speaking, we can first distinguish between data-independent and data-dependent types of chance corrections. Within the latter category, we can further subdivide measures by whether they rely on information from the entire sample to calculate the chance correction or only from the specific dyad. Figure 3 shows the resulting classification tree.

Currently, the most prominent agreement index in international relations research is Signorino and Ritter's (1999)  $S$ . In its simplest and most widely used form, this index is given by  $s = 1 - 2 * \sum_{l=1}^r \frac{|y_l - x_l|}{d_{\max}}$ , where  $y_l$  and  $x_l$  stand for the type of vote countries  $Y$  and  $X$  cast on resolution  $l$ ,  $d_{\max}$  for the theoretically possible maximum distance between  $y$  and  $x$  values, and the summation is over all resolutions  $l = 1, \dots, r$ . Thus, for each resolution,  $S$  first calculates the



**Figure 3.** Classification of chance-correction approaches.

distance between the two countries' vote variable values and then normalizes the observed distance by dividing it by the theoretically possible maximum distance. These normalized distance values are then summed up over all resolutions. Translated into our notation, the sum of normalized observed distances in  $S$  corresponds to the proportion of disagreement derived from a contingency table:

$$\begin{aligned}
 s &= 1 - 2 * \sum_{l=1}^r \frac{|y_l - x_l|}{d_{\max}} = 1 - 2 * \sum_{i=1}^k \sum_{j=1}^k \frac{w_{ij} f_{ij}}{w_{\max} f_{ij}} = 1 - 2 * \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} f_{ij}}{w_{\max} f_{..}} \\
 &= 1 - 2 * \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{ij}}{w_{\max}} = 1 - 2 * D_0 = 1 - \frac{D_0}{0.5}
 \end{aligned}
 \tag{4}$$

The reformulation makes it clear that  $S$  is simply a linear function of the proportion of disagreement  $D_0$ . The multiplication by 2 “stretches” the disagreement values from its original range between 0 and 1 to a range between 0 and 2. The subtraction of the resulting value from 1 reverses the polarity of the measure and rescales it to a range between  $-1$  indicating complete disagreement and 1 indicating complete agreement. The equation for  $S$  can be further reformulated to bring it completely in line with the format of the general equation for chance-corrected agreement indices. Rather than multiplying the observed proportion of disagreement by 2, we can equivalently divide it by 0.5. Thus, when interpreted as a chance-corrected agreement index, the expected proportion of disagreement of  $S$  is 0.5. In other words, half of the theoretically possible maximum proportion of disagreement is expected to occur by chance. In general, disagreement expected by chance is given by the following formula for all chance-corrected agreement indices:



$$D_e = \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} m_i \cdot m_j}{w_{\max}} \quad (5)$$

Different indices vary only in the assumptions they make about the marginals  $m_i$  and  $m_j$  of the vote variables used to calculate the expected disagreement. In other words, they differ only in their assumptions about states' propensities to vote a certain way (see Table A12 in Web appendix for a summary of these assumptions).

Table 4 illustrates how the disagreement expected by chance differs depending on these assumptions and how the different chance-corrections then lead to different similarity values. In the case of *S*, the marginals for the calculation of the expected disagreement are not related to the observed contingency table. Therefore, *S* implicitly relies on a data-independent chance-correction based on an expected disagreement score of 0.5. An expected disagreement by chance of 0.5 can be generated through various combinations of marginal distributions, including any that involves one member state having a 0.5 propensity to fall into each of the extreme categories (that is, *yea* or *nay*) and a zero propensity to fall into the intermediate category (that is, *abstain*). However, if we assume that both member states have the same propensities to vote in a certain way, that is, assume that their marginal distributions are identical, only the situation in which both member states have a 0.5 propensity to vote *yea* and *nay* and a zero propensity to abstain produces an expected disagreement of 0.5. The contingency table of expected proportions generated by these marginals, together with the relevant disagreement weights, is depicted in Panel B of Table 4.

The assumptions about the form of the marginal distributions used to calculate the chance correction of *S* are hard to justify on substantive grounds.<sup>20</sup> Assuming that states have a 50% probability of voting *yea* or *nay* and a 0% probability of abstaining contradicts both common sense and available empirical information.<sup>21</sup> A somewhat more plausible, also data-independent way of correcting for chance is to assume that states have the same propensity of 1/3 to vote either *yea*, *nay*, or *abstain* (for example, Lijphart 1963:906–908; Mokken and Stokman 1985:186–187). Panel C of Table 4 illustrates the case where chance disagreement is calculated based on such uniform marginals. Note that the chance disagreement based on uniform marginals is smaller than the chance disagreement implicitly assumed by *S*. Indeed, Mokken and Stokman (1985:187) assert that the assumption about the extreme bimodal marginal distribution

<sup>20</sup>Mokken and Stokman (1985:187–188) argue that this chance correction is useful for measuring the cohesion of a decision-making body as a whole.

<sup>21</sup>The lack of plausible assumptions about the marginal distributions used in the calculation of chance disagreement in *S* is understandable, given that the correction for chance disagreement was not an explicit goal in the development of this measure.

**Table 4.** Calculation of Indices Based on Different Assumptions about Marginals.

A. Observed disagreement						B. Signorino and Ritter's $S$					
USA						USA					
	1	2	3			1	2	3			
	1	0.00	0.10	0.20	0.30		1	0.25	0.00	0.25	0.50
	0		1	2			0	1	2		
USSR	2	0.00	0.10	0.10	0.20	USSR	2	0.00	0.00	0.00	0.00
	1		0	1			1	0	1		
	3	0.10	0.00	0.40	0.50		3	0.25	0.00	0.25	0.50
	2		1	0			2	1	0		
		0.10	0.20	0.70	1			0.50	0.00	0.50	1
$D_0 = (4 * 0.10 + 2 * 0.20)/2 = 0.40$						$D_e^s = (4 * 0.25)/2 = 0.50$					
						$A_e^s = 1 - \frac{D_0}{D_e^s} = 1 - \frac{0.40}{0.50} = 0.2$					
C. Uniform marginals						D. Country/resolution average marginals					
USA						USA					
	1	2	3			1	2	3			
	1	0.11	0.11	0.11	0.33		1	0.03	0.05	0.10	0.18
	0		1	2			0	1	2		
USSR	2	0.11	0.11	0.11	0.33	USSR	2	0.05	0.08	0.15	0.28
	1		0	1			1	0	1		
	3	0.11	0.11	0.11	0.33		3	0.10	0.15	0.29	0.54
	2		1	0			2	1	0		
		0.33	0.33	0.33	1			0.18	0.28	0.54	1
$D_e^u = (8 * 0.11)/2 = 0.44$						$D_e^v = (2 * 0.05 + 4 * 0.10 + 2 * 0.15)/2 = 0.40$					
$A_e^u = 1 - \frac{D_0}{D_e^u} = 1 - \frac{0.40}{0.44} = 0.10$						$A_e^k = 1 - \frac{D_0}{D_e^k} = 1 - \frac{0.40}{0.40} = 0.00$					
E. Scott's $\tau$						F. Cohen's $\kappa$					
USA						USA					
	1	2	3			1	2	3			
	1	0.04	0.04	0.12	0.20		1	0.03	0.06	0.21	0.30
	0		1	2			0	1	2		
USSR	2	0.04	0.04	0.12	0.20	USSR	2	0.02	0.04	0.14	0.20
	1		0	1			1	0	1		
	3	0.12	0.12	0.36	0.60		3	0.05	0.10	0.35	0.50
	2		1	0			2	1	0		
		0.20	0.20	0.60	1			0.10	0.20	0.70	1
$D_e^\pi = (2 * 0.04 + 6 * 0.12)/2 = 0.40.$						$D_e^\pi = \left( \frac{0.06 + 2 * 0.21 + 0.14 +}{0.02 + 2 * 0.05 + 0.10} \right) / 2 = 0.84$					
$A_e^\pi = 1 - \frac{D_0}{D_e^\pi} = 1 - \frac{0.40}{0.40} = 0.00.$						$A_e^\pi = 1 - \frac{D_0}{D_e^\pi} = 1 - \frac{0.40}{0.42} = 0.05.$					

used to calculate the expected disagreement for  $S$  yields the theoretically possible maximum expected disagreement. This assertion seems to only hold for indices that assume that the marginal distributions are symmetrical (that is, identical for both states).<sup>22</sup> With the exception of Cohen's  $\kappa$ , all of the indices discussed here make this assumption.

<sup>22</sup>It is easy to construct an example of a contingency table with asymmetric marginal distributions that yields a higher expected proportion of disagreement value than 0.5.

Just like any data-independent approach to specifying the marginal distributions, the choice of uniform values might be criticized for neglecting empirical information about the actual voting behavior. As already recognized by Lijphart (1963:906), few resolutions are put to a vote in the UNGA that do not pass, and abstentions are rarer than either *yea* or *nay* votes. Thus, rather than equal probabilities, he suggests it is more plausible to assume that the probability of voting *yea* is higher than the probability of voting *nay*, and that the latter in turn is higher than the probability of abstaining. Rather than assuming relatively arbitrary values for the marginal distributions based on rules of thumb derived from general voting patterns in the UNGA, Mokken and Stokman (1985:187) go one step further in suggesting that these values could be directly estimated from the information in the sample. They propose to estimate the marginals by computing, for each resolution, the proportion of states voting in favor, against, and abstaining. Subsequently, the proportions are averaged over all resolutions adopted during the particular session or time period. We call this approach “resolution average marginals,” as proportions of states voting in a certain way on a particular resolution are averaged over all resolutions to estimate the marginals (see Panel D in Table 4).

The “country average marginals” approach is similar, but here the vote proportions are first calculated for individual states across all resolutions and then averaged over all states. When there are no missing values in the voting matrix, as in the example of Table 4, the two approaches yield identical results. However, in real-world UNGA voting, the voting matrix often has missing values because some member states might not have been members of the UN for the entirety of the particular time period for which the agreement index is being calculated, or they have not been taking part in one or more of the votes for other unknown reasons. In light of missing values, the sequence in which vote proportions and averages are being calculated to estimate the marginal distributions matters. Given the nonuniform shape of actually observed marginal distributions, these empirically informed chance-correction approaches are certainly an improvement over data-independent approaches, especially when a large number of consensus votes are part of the sample. The chance-corrected approaches provide measures of dyadic agreement over and above the agreement expected for a dyad with average sample marginal. In the face of many consensus votes, expected agreement will be higher, lowering the value of the similarity variable. If we have reasons to believe that consensus votes are somewhat less indicative of true agreement than recorded *yea* votes, this effect is certainly desirable. Also, changes in the average marginal distributions over time are likely to be mostly the result of agenda effects. When more resolutions with “agreeable” content are being tabled in a certain year, the rate of consensus decisions will be generally higher compared to the situation where mostly resolutions with controversial

content are being tabled in a certain year, even if the underlying policy positions of states have remained constant. Putting the observed agreement into relation to the agreement expected based on average marginals adjusts the similarity values for those types of agenda effects.

However, the effect of changes in extrinsic variables, such as the agenda, on states' vote marginals might not be uniform across all states. Both the voting behavior of particular dyads and individual countries within dyads might be unduly affected by changes in those variables as well. Scott's (1955)  $\pi$  and Cohen's (1968)  $\kappa$  address these issues. The country average marginals approach is basically an extension of the chance-correction approach used in the calculation of Scott's  $\pi$ . While the country average marginals approach averages the propensities of states to vote in a certain way over all states in the sample, Scott's  $\pi$  only averages the vote propensities of the two states that form part of the particular dyad (see panel E of Table 4). In this respect, Scott's  $\pi$  is more flexible and able to not only adjust for factors that affect the voting behavior of all states in the sample equally (for example, consensus votes) but also for factors that affect only the voting behavior of the particular dyad members in the same way. For example, if over time more resolutions are put on the agenda about issues on which both dyad members generally agree, raw similarity scores increase even if the underlying policy preferences remain stable. If this type of agenda effect is relatively unique to particular dyads or affects dyads in different ways, chance-corrections based on the marginal distributions of the entire sample of dyads cannot alleviate the problem. Only dyad-specific chance-corrections limit this type of distortion.

Yet Scott's  $\pi$  still assumes that both dyad members have identical propensities to vote in a certain way, although good reasons exist to expect that certain factors have divergent effects on the voting behavior of dyad members. In general, expected agreement based on the average dyad marginal will be larger if the marginal distributions are symmetrical rather than asymmetrical. In other words, Scott's  $\pi$  takes a lower value if dyad members differ in their propensities to vote in a certain way than if they share the same propensities. At first sight, this seems reasonable. Differences in the marginal distributions show up as differences in the agreement score.<sup>23</sup> However, a case can be made that the number of times a state votes in a certain way depends mainly on the types of issues decided upon in the time period under investigation. For example, during a period where many resolutions on the Middle East conflict are being adopted, the United States might vote against France quite a large number of times. In other

<sup>23</sup>Häge (2011:293) makes the case that the assumptions of Scott's  $\pi$  are more appropriate for measuring foreign policy similarity based on UNGA voting data. In terms of the relatively low costs of creating a UNGA voting tie compared to an alliance tie, this makes sense. However, in the case of UNGA data, the main reason why individual states may systematically differ in their propensity to vote in a certain way has less to do with differential costs, given that voting is relatively "cheap" regardless of what type of vote is being cast (see Hovet 1960) but with the content of the agenda they are asked to vote upon.

periods, where the Middle East conflict is less salient, the number of opposing votes might be much smaller in that dyad even though no change in the underlying foreign policy preferences has occurred. In general, the same agenda change might lead to more covoting for some dyads, while resulting in more opposing votes for others. Scott's  $\pi$  only accounts for the first possibility. In contrast, Cohen's  $\kappa$  allows each dyad member to have its own independent marginal distribution for the calculation of the proportion of expected agreement (see panel F of Table 4). The measure directly uses the marginal distributions of the observed contingency table to estimate the expected marginal distributions. In comparison to Scott's  $\pi$ , Cohen's  $\kappa$  results in a lower expected agreement value if marginal distributions are asymmetrical, adjusting similarity values upwards. Given that Cohen's  $\kappa$  is most versatile in adjusting for both the inclusion of consensus votes and the potentially divergent effects on voting behavior resulting from changes in the agenda and other factors, the following replication studies focus on the performance of this chance-corrected agreement index compared to the widely used  $S$  proposed by Signorino and Ritter's (1999).<sup>24</sup>

### Replication of Alesina and Dollar (2000)

In his study on chance-corrected agreement indices, Häge (2011) demonstrates that  $S$  and chance-corrected agreement indices like Cohen's  $\kappa$  and Scott's  $\pi$  are not interchangeable and can lead to very different conclusions drawn from statistical analyses. In a replication of Gartzke's (2007) study of the determinants of interstate war onset, he shows that the results are only consistent with Gartzke's theoretical claims once  $S$  is replaced by  $\kappa$  or  $\pi$  in the regression model.

Instead of drawing on the same example, we turn to another literature in which affinity and similarity measures are in frequent use, namely, the literature on foreign aid. In a pathbreaking study, Alesina and Dollar (2000) find that political and strategic reasons explain to a significant part aid allocation both generally and by individual countries like the United States. In what follows, we carry out replications of two models of Alesina and Dollar's (2000) study on total bilateral aid and US bilateral aid given to recipient countries in 5-year periods.<sup>25</sup> These models, apart from economic and social explanatory variables, also comprise political factors such as civil liberties and measures of whether a recipient country was a friend of a specific donor country. The latter measure is operationalized as the proportion of votes in the UNGA in which the two countries were in agreement.<sup>26</sup>

<sup>24</sup>In the Web appendix we also report replication results based on the other four similarity measures discussed earlier.

<sup>25</sup>We obtained the replication data from the AidData website (<http://aiddata.org/content/index/Research/replication-datasets>), and David Dollar provided greatly appreciated help in using it.

For this replication, we rely on the Alesina and Dollar (2000) data and complement it with our own similarity measures based on new data of UNGA voting. Most studies rely on Voeten's (2000) UNGA voting data, which relies in part on Gartzke's (1998), in part on Kim and Russett's (1996) and on Alker and Russett's (1965) data (see also Strezhnev and Voeten 2012). Unfortunately, combining data from different sources has led to a situation in which the inclusion criteria vary across time periods (for example, votes on amendments, etc., are included until the 1970s but figure no longer in the data for more recent periods). For this reason we rely on Hug's (2012) data, which comprise, based on a common source, all votes on resolutions as well as information on all resolutions debated in the UNGA. As we have information on both resolutions adopted through roll-call votes and resolutions adopted through consensus votes, we proceed as follows:

First, we generate for each year a data set that only comprises the member state voting records on resolutions adopted through roll-call votes.

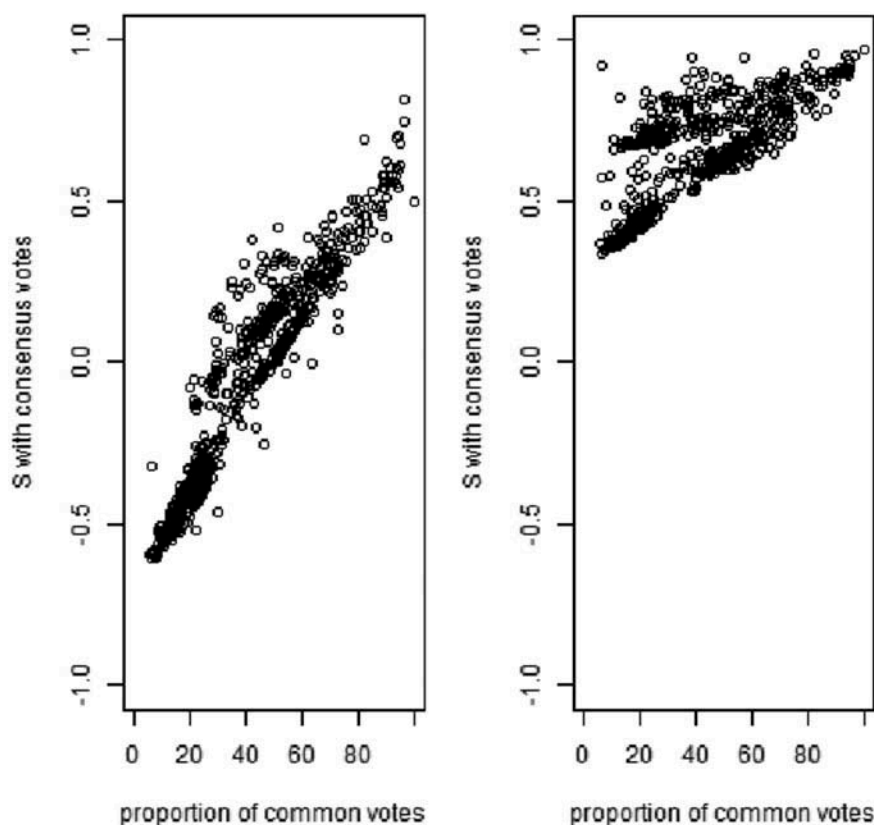
Second, we generate an imputed data set where for all states that were members of the UN at the time of the vote, we assume that they voted in favor of all resolutions adopted without a vote.<sup>27</sup>

As Alesina and Dollar's (2000) study uses 5-year periods as the temporal unit of analysis, we followed their approach used for all other variables and aggregated our yearly similarity measures based on our imputed UNGA voting data by calculating 5-year averages. We then merged our data with Alesina and Dollar's (2000) replication data set. As a first step in the analysis, this allows us to compare our similarity measures with those employed in the original study, namely, the proportion of common votes between the aid recipient and the United States (and other countries). Figures 4 and 5 depict the relationships between the proportion of common votes and Signorino and Ritter's (1999)  $S$  and Cohen's  $\kappa$  respectively. The left panel of each figure provides the similarity values based on roll-call votes only, while the right panel provides similarity values that also take consensus votes into account.

In Figure 4, where we compare  $S$  to the proportion of common votes, we find that in the left panel without consensus votes, the two measures are closely related. Given that  $S$  is a linear transformation of the proportion of common votes, this is not surprising. Indeed, any deviation from a perfect relationship between the two variables must be due to differences in the underlying data. When taking consensus votes into account (right panel in Figure 4), we find generally much higher  $S$  values but also a much weaker relationship between  $S$  and the proportion of common votes.

In Figure 5, where we rely on Cohen's  $\kappa$ , already the left panel omitting consensus votes shows a rather weak relationship between the values of  $\kappa$  and

<sup>27</sup> Again, it is important to note that we make the assumption that adoptions without a vote signal unanimous support for the resolution in question. As noted in footnote 8, we omit nonrecorded votes for which only the marginal distribution is recorded. Results based on imputed data sets taking those nonrecorded votes into account are reported in the Web appendix. All the data will be made available on dataverse upon publication.



**Figure 4.** Ritter and Signorino's  $S$  vs. proportion of common votes.

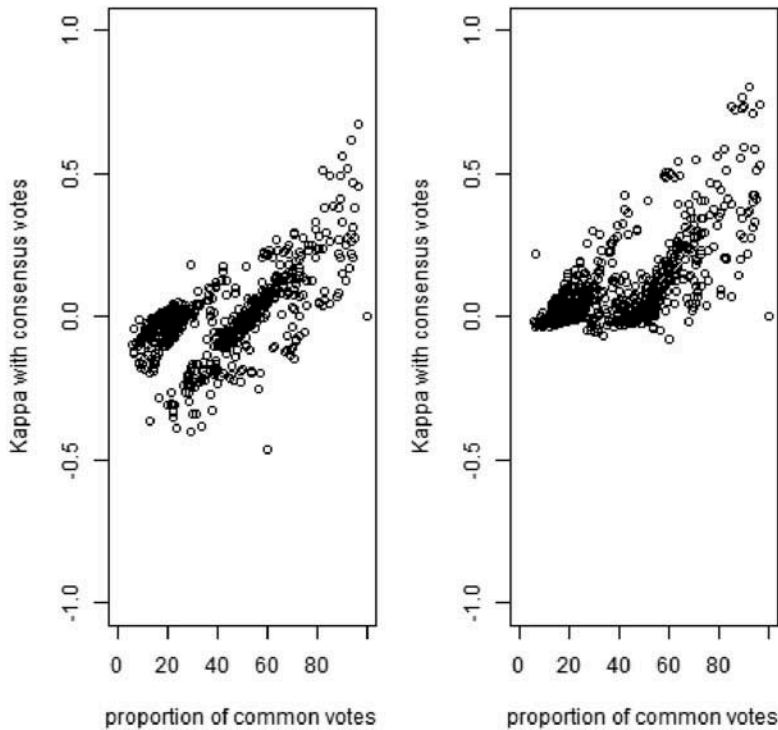
*Note.* The figure plots similarity values averaged over 5-year periods for dyads including the US as donor country.

the proportion of common votes. Again, once we include consensus votes, the value of  $\kappa$  generally increases, and the relationship with Alesina and Dollar's (2000) proportion of common votes becomes considerably more blurred. Hence, it is likely that the proportion of common votes, by not considering consensus votes, is actually measuring something quite distinct from affinity.

We assess the effects of these different measurement strategies on the conclusions of Alesina and Dollar's (2000) analyses by reestimating one of their models, focusing on bilateral aid obtained from the US (Table 5).<sup>28</sup> While Alesina and Dollar (2000) make their data available, there are very few indications on how these data were used to produce the results reported in their paper. Thus, we first report in the first column the results reported in Alesina and Dollar's (2000) article before showing our replication in the other columns. We then replace in these models the proportion of common votes between the aid recipient and the US (or

<sup>28</sup>In the Web appendix we also report replications of a model by Alesina and Dollar (2000) focusing on total bilateral aid. For the model reported in Table 3, we list in the Web appendix (Table A11) the list of countries covered and the number of cases.





**Figure 5.** Cohen's  $\kappa$  vs. proportion of common votes.

*Note.* The figure plots similarity values averaged over 5-year periods for dyads including the US as donor country.

Japan respectively) with  $S$  and  $\kappa$ . In the first two models, the affinity measures are based only on roll-call votes, in the last two models we also include information on consensus votes in the calculation of  $S$  and  $\kappa$ .

In Table 5 we report the results of our replication that focuses on explaining US bilateral aid. We are unable to reproduce the positive effect of GDP per capita reported by Alesina and Dollar (2000).<sup>29</sup> For the other variables, we are able to approximate the original results, except that no former colony of the US has nonmissing data on all variables, which is the reason why this variable drops from our replication. We are only able to partly replicate the positive effect of voting with the US on obtaining aid from this country. When we consider  $S$  and  $\kappa$  as similarity measures while ignoring consensus votes, we obtain contrasting results. While for  $\kappa$  we find the positive effect found by Alesina and Dollar (2000), for  $S$  the effect is negative, but both effects are not

<sup>29</sup>Given the robustness of the negative effect of this variable in the remaining models in Table 3, we can only suspect a typo in Alesina and Dollar's (2000) article. Regarding Alesina and Dollar's (2000) model, one might also suspect that changes in the dependent variable over time are not only affected by their independent variables but also by past aid allocations (we thank an anonymous reviewer for alerting us to this point). As the goal of our replication analysis is to show the sensitivity of Alesina and Dollar's (2000) result to changes in the measures of similarity, it seems inappropriate to change the underlying empirical model.

**Table 5.** Replication of Alesina and Dollar (2000), Bilateral Aid by US (Linear Regression with White Robust Standard Errors).

	Similarity Measure					
	Without Consensus Votes				With Consensus Votes	
	Proportion of Agreement	Chance Corrected			Chance Corrected	
		S	$\kappa$		S	$\kappa$
	b	b	b	b	b	b
	(t)	(SE)	(SE)	(SE)	(SE)	(SE)
Log GDP per capita	1.840*	-1.662*	-1.768*	-1.736*	-1.714*	-1.689*
	—	(0.153)	(0.155)	(0.154)	(0.154)	(0.154)
Economic openness	1.300*	0.818*	1.048*	1.013*	0.964*	0.956*
	(4.02)	(0.284)	(0.288)	(0.288)	(0.288)	(0.286)
democracy	0.570*	0.388*	0.445*	0.420*	0.404*	0.386*
	(8.07)	(0.064)	(0.065)	(0.064)	(0.065)	(0.065)
Friend of USA (UNGA voting)	0.060*	0.042*	-3.102	0.818	1.922	3.659*
	(3.60)	(0.014)	(1.996)	(1.052)	(1.016)	(1.406)
Log years as colony of US	0.39*					
	(1.69)					
Log years as colony not of US	0.08	-0.007	-0.010*	-0.009*	-0.009*	-0.008*
	(1.33)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Egypt	40.090*	4.514*	4.545*	4.554*	4.528*	4.509*
	(4.14)	(0.893)	(0.896)	(0.899)	(0.895)	(0.891)
Israel	5.040*	4.759*	8.065*	6.709*	5.804*	5.484*
	(3.94)	(1.112)	(1.002)	(1.070)	(1.125)	(1.070)
Percent Muslims	0.010*	0.023*	0.025*	0.025*	0.025*	0.025*
	(1.98)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)
Percent Catholics	0.010	0.018*	0.023*	0.021*	0.020*	0.020*
	(1.69)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Percent other religions (Hindu)	0.004	0.011	0.011	0.012	0.013	0.014*
	(0.05)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
1970–1974		9.438*	14.128*	11.927*	11.574*	11.716*
		(1.327)	(1.736)	(1.096)	(1.110)	(1.085)
1975–1979		9.340*	14.113*	11.850*	11.706*	11.691*
		(1.339)	(1.788)	(1.109)	(1.106)	(1.098)
1980–1984		11.019*	13.744*	12.176*	12.815*	12.075*
		(1.152)	(1.480)	(1.115)	(1.152)	(1.105)
1985–1989		11.170*	13.443*	12.037*	12.892*	11.949*
		(1.121)	(1.414)	(1.106)	(1.184)	(1.097)
1990–1994		10.680*	14.192*	11.853*	12.545*	11.839*
		(1.152)	(1.844)	(1.108)	(1.154)	(1.097)
N	364	364	358	358	358	358
R <sup>2</sup>	0.5	0.705	0.704	0.702	0.705	0.708
Resid. s.d.		1.722	1.728	1.732	1.725	1.717

Note. Robust standard errors in parentheses, except in the first column where *t* values, based on robust standard errors, are reported. \* indicates significance at  $p < .05$ .

statistically significant. When considering consensus votes in the calculation of those similarity measures as well, we find two positive coefficients, but only the effect of  $\kappa$  remains statistically significant.<sup>30</sup> Consequently, if Alesina and Dollar (2000) had used the currently predominant measure of preference similarity in their analysis, their conclusion would have been that political

<sup>30</sup>When replicating these analyses and also taking into account nonrecorded votes through imputed data sets, we find the same pattern of coefficients (see Table A6 in the Web appendix).

and strategic explanations are unimportant for explaining US bilateral aid. Only when using both consensus votes and a chance correction, voting similarity as measured by  $\kappa$  appears to significantly affect US bilateral aid.

## Conclusion

An increasing number of studies dealing with a variety of topics relies on similarity measures based on voting records in the UNGA to measure preferences of governments. As several studies have shown, the most widely used measures have considerable shortcomings. First, as illustrated by Häge (2011), chance agreement is not adjusted for in an explicit and sensible way by most commonly used measures. Second, Bailey et al. (2013) convincingly highlight that these same measures suffer from agenda effects as resolutions often deal with very topical issues on conflict. Finally, we stressed in this article that neglecting the varying share of consensus votes is equally likely to lead to biases in these measures.

We first demonstrated this problem based on “artificial data,” showing that neglecting consensus votes is likely to underestimate affinities among country pairs. Under the assumption that resolutions that are adopted without a vote have the tacit support of all UNGA members at the time of the vote, we generated a data set comprising information on all resolutions adopted both with and without an explicit vote. Not surprisingly, when compared to traditional measures like the proportion of common votes (leaving aside consensus votes), measures considering consensus votes as well show higher levels of affinity (and thus also less variation). When replicating Alesina and Dollar’s (2000) influential study on the political and strategic determinants of bilateral aid, we find that many of their main findings are not robust to the inclusion of consensus votes. More specifically, their effect of preference similarities with the United States on US bilateral aid can only be reproduced if consensus votes are integrated and a chance-corrected measure like  $\kappa$  is used. In the absence of this, we find no effect of preference similarity on US bilateral aid. Conversely, using this same measure with consensus votes preference similarity with Japan, contrary to Alesina and Dollar’s (2000) result, fails to affect the level of overall bilateral aid. In addition, we can show that even when calculating similarity measures only based on UNGA decisions deemed important by the US State Department (see Thacker 1999), the results reported by Alesina and Dollar (2000) fail to be robust.<sup>31</sup> Consequently, we find very little evidence, if any, supporting the claim that voting in the UNGA affects aid allocation.

Hence, scholars wishing to use measures of affinity and similarity should be prudent when relying on existing measures. The latter do not control for possible chance agreements and by neglecting consensus votes introduce biases in their estimates. These biases, as we have demonstrated in a replication study, can also have considerable substantive consequences. Our approach,

---

<sup>31</sup>The latter two results derive from analyses reported in the Web appendix.

however, does not deal directly and explicitly with the problem highlighted by Bailey et al. (2013), namely, possible agenda effects. As we noted, both Scott's  $\pi$  and Cohen's  $\kappa$ , by relying on data-dependent chance-correction, may implicitly adjust for such effects. Bailey et al.'s (2013) approach to solve the problem of agenda effects can by definition not consider consensus votes and is thus likely to lead to biased estimates. This may occur if resolutions that prove very important for estimating the ideal-point of UNGA member states are adopted by consensus or through unrecorded votes. Recently, Marbach (2015) has proposed an innovative way of integrating unrecorded votes into an empirical model that might be extended towards an IRT model. Consequently, future research has to show whether further-developed IRT models may take into account both agenda effects and unrecorded votes and whether their estimates for similarities improve upon the measures proposed in this article.

## References

- Alesina, Alberto, and David Dollar. (2000) Who Gives Foreign Aid to Whom and Why? *Journal of Economic Growth* 5(1):33–63.
- Alker, Hayward R., and Bruce Russett. (1965) *World Politics in the General Assembly*. New Haven, CT: Yale University Press.
- Bailey, Michael A., Anton Strezhnev, and Erik Voeten. (2013) Estimating State Preferences from United Nations Data. Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, April 11–14.
- Blake, Daniel J., and Autumn Lockwood Payton. (2015) Balancing Design Objectives: Analyzing New Data on Voting Rules in Intergovernmental Organizations. *Review of International Organizations* 10(3):377–402.
- Cohen, Jacob. (1968) Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70:213–20.
- 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):1–27.
- Gartzke, Erik. (2007) The Capitalist Peace. *American Journal of Political Science* 51:166–191.
- Häge, Frank M. (2011) Choice or Circumstance? Adjusting Measures of Foreign Policy Similarity for Chance Agreement. *Political Analysis* 19(3):287–305.
- Häge, Frank M. (2013) Coalition-Building and Consensus in the Council of the European Union. *British Journal of Political Science* 43(3):481–504.
- Hovet, Thomas. (1960). *Bloc Politics in the United Nations*. Cambridge, MA: Harvard University Press.
- Hug, Simon. (2010) Selection Effects in Roll Call Votes. *British Journal of Political Science* 40 (1):225–235.
- Hug, Simon. (2012) What's in a Vote? Paper presented at the 5th Conference on The Political Economy of International Organizations, Villanova, January 26–28.
- Jessee, Stephen A. (2010) Issues in Scaling Citizens and Legislator Ideology Together. Working paper, The University of Texas at Austin.
- Kim, Soo Yeon, and Bruce Russett. (1996) The New Politics of Voting Alignments in the United Nations General Assembly. *International Organization* 50(4):629–652.

- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. (2001) Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 95(1):49–69.
- Lewis, Jeffrey B., and Chris Tausanovitch. (2013) Has Joint Scaling Solved the Achen Objective to Miller and Stokes? Working paper, UCLA Department of Political Science.
- Lijphart, Arend. (1963) The Analysis of Bloc Voting in the General Assembly: A Critique and a Proposal. *American Political Science Review* 57(4):902–917.
- Marbach, Moritz. (2015) A Discrete Choice Model for the Analysis of Aggregated Roll Call Voting Records. Paper prepared for presentation at the 8th Annual Conference on The Political Economy of International Organizations, Berlin, February 12–14.
- Mattes, Michaela, Brett Ashley Leeds, and Royce Carroll. (2015) Leadership Turnover and Foreign Policy Change: Societal Interests, Domestic Institutions, and Voting in the United Nations. *International Studies Quarterly* 59(2):280–290.
- Mokken, Robert Jan, and Frans N. Stokman. (1985) Legislative Analysis: Methodology for the Analysis of Groups and Coalitions. In *Coalition Formation*, edited by Henk A. M. Wilke. Amsterdam: North-Holland, pp. 173–227.
- Peterson, M. J. (2005) *The United Nations General Assembly*. Abingdon, Oxon: Routledge.
- Scott, William A. (1955) Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly* 19(3):321–5.
- Signorino, Curtis S., and Jeffrey M. Ritter. (1999) Tau-b or Not Tau-b: Measuring the Similarity of Foreign Policy Positions. *International Studies Quarterly* 43:115–144.
- Stokman, Frans N. (1977) *Roll Calls and Sponsorship: A Methodological Analysis of Third World Group Formation in the United Nations*. Leyden: A. W. Sijthoff.
- Strezhnev, Anton, and Erik Voeten. (2012) United Nations General Assembly Voting Data. Available at <http://hdl.handle.net/1902.1/12379> UNF:5:iiB+pKXYsW9xMMP2wfY1oQ==V3 [Version].
- Thacker, Strom. (1999) The High Politics of IMF Lending. *World Politics* 52(1):38–75.
- Voeten, Erik. (2000) Clashes in the Assembly. *International Organization* 54(2):185–215.