Consensus voting and similarity measures in IOs[1]

Frank M. Häge[2] and Simon Hug[3]

Department of Politics and Public Administration, University of Limerick

and

Département de science politique et relations internationales, Université de Genève

Paper prepared for a possible presentation at the 2014 PEIO conference

(January 13-15, 2014 Princeton)

First version: February 2013, this version: January 7, 2014

Abstract

Voting behavior in international organizations, most notably in the United Nations General Assembly (UNGA), is often used to infer the similarity of foreign policy preferences of member states. Most of these measures ignore, however, that particular co-voting patterns may appear simply by chance (Häge 2011) and that these patterns of agreement (or the absence thereof) are only observable if decisions are reached through recorded votes. As the frequency of such roll-call votes changes considerably in most international organizations and particularly in the UNGA over time, frequently used similarity and affinity measures offer a misleading picture. Based on a complete data set of UNGA resolution-related decisions, we demonstrate how taking different forms of chance agreement and the relative prevalence of consensus decisions into account affects conclusions about the similarity of member states' foreign policy positions.

[1] An earlier version of this paper was presented at the 2013 EPSA Annual Meeting (June 20-22, 2013 Barcelona). We wish to thank the discussant of the panel Kris Ramsey, as well as other participants for very helpful comments. Simon gratefully acknowledges the research assistance by Simone Wegmann and Reto Wüest and partial financial support by the Swiss National Science Foundation (Grant-No 100012-129737).

[2] Department of Politics and Public Administration, University of Limerick, Limerick, Ireland; phone: +353-61-23-4897; email: frank.haege@ul.ie.

[3] Département de science politique et relations internationales, Faculté des sciences économiques et sociales; Université de Genève; 40 Bd du Pont d'Arve; 1211 Genève 4; Switzerland; phone +41-22-379-83-78; email: simon.hug@unige.ch.

1 Introduction

Affinity measures based on voting in the United Nations General Assembly (UNGA) have experienced an increasing popularity. In a recent paper, Bailey, Strezhnev and Voeten (2013) mention that since Gartzke's (1998) prominent use of such data, almost 100 articles and papers have relied on voting data to construct preference measures for states and their governments (for a more general survey article on voting data in the UNGA, see Voeten 2013, 55, who mentions 50 such studies).

These affinity measures are all predicated on the idea that observing a pair of countries voting frequently in unison is the result of preference affinities (see, for instance, Alesina & Dollar 2000). In the context of voting in the UNGA, however, such measures are problematic for at least three different reasons. First, as Häge (2011) argues, many of these measures do not take into account the possibility of chance agreement, which is linked to specific alliance patterns (for a related argument, see Stokman 1977 and Mokken & Stokman 1985). Thus, he proposes affinity measures that correct for chance agreement.

Second, Bailey, Strezhnev & Voeten (2013) convincingly show that affinity measures cannot address the issue of changing agendas. More specifically, if due to a particular conflict a series of resolutions are voted upon, the preference configuration related to this conflict will strongly affect affinity measures. According to these authors, a more appropriate item-response theory (IRT) model with bridging observations across sessions formed by resolutions with very similar contents, allows to circumvent this problem.

A third issue, however, has so far remained largely unaddressed, namely the fact that consensus voting plays an important role in many international organizations in general and the UNGA in particular. In the latter, for instance, only a small share of resolutions are actually voted upon, while a large majority is adopted without a vote (i.e., consensus vote).[4] Existing affinity measures and IRT-models rely exclusively on data about contested votes. Resolutions adopted without a vote are not reflected in these measures. As the share of resolutions adopted without a vote varies across years in the UNGA (and also across issue domains, see Hug 2012, Skougarevskiy 2012) both affinity measures and estimates from IRT-models are likely to be affected by these missing "votes."

[4] In all of the paper we will use adoption without votes as synonym of consensus vote (as does, implicitly, much of the literature, see Cassan 1977, Blake and Lockwood Payton 2009 and Lockwood Payton 2010, 2011).

In the present paper, we address this issue and show how it may be addressed in the context of studies using affinity scores.[5] We find that neglecting consensus votes when using UNGA data may seriously affect inferences. More specifically, we replicate the study by Alesina and Dollar (2000) on the political and strategic elements explaining why aid recipients obtain bilateral aid from specific donors. We find that political closeness as measured on the basis of UNGA votes loses most of its importance in explaining aid allocation once we account for chance agreement and include information on consensus votes.

In the next section we briefy discuss the role of consensus voting in international organizations. Section three presents a brief overview of research using affinity measures based on UNGA voting data. It also highlights how the practice of consensus voting might affect the results offered in these studies. In section four we demonstrate in detail how chance agreements and consensus votes (and their neglect) affect similarity measures. Section five presents a new data set on UNGA voting comprising, for the first time, information about resolutions adopted without a vote. In a replication of Alesina & Dollar's (2000) study, this section shows that taking consensus votes into account considerably affects findings about the relationship between political closeness and bilateral aid. We then conclude in section six.

2 Consensus voting in international organizations

In numerous bodies of international organizations decisions are reached by voting. As Blake and Lockwood Payton (2009) nicely show, the exact rules for decision-making differ considerably across these various international bodies. In many of them, also in those of the United Nation (see Cassan 1977, Abi-Saab 1997), consensus decisions are of considerable importance (see for a discussion and explanation of voting rules, Lockwood Payton 2010, 2011). The UNGA, for instance, describes their voting practices in the following way (http://www.un.org/Depts/dhl/resguide/gavote.htm, accessed September 7, 2011):

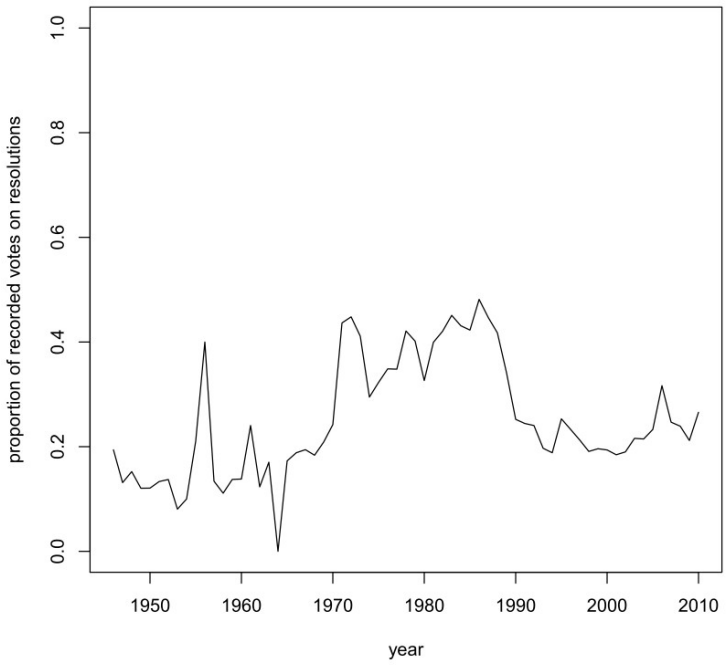
> "The majority of General Assembly resolutions are adopted without a vote. If a vote is taken, it can be documented in two ways: either as a recorded vote or as a summary of the result. Only a recorded vote, which must be requested before the voting is conducted, will clearly identify the stand that a Member State took on the issue under discussion. If such a request is not put forth,

[5] 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.

only the voting summary (i.e., the number of countries which voted for or against a resolution as well as those who abstained) will be made available, without identification of how an individual Member State voted."

While the large share of resolutions being adopted without a vote (i.e., consensus voting) is acknowledged in the literature, its variation over time has been largely ignored.[6] Thus, Figure 1 depicts the share of recorded votes on final passage of resolutions in the UNGA in the period between 1945 and 2011 (*Source*: Hug 2012). The figure clearly shows that the share of recorded votes (with the exception of 1964) has varied between a low of approximately 10 percent and a high of almost 50 percent. This implies that focusing only on recorded votes leaves aside between 50 and 90 percent of all decisions on resolutions in the UNGA.[7]

Figure 1: Proportion of recorded votes on resolutions in the UNGA over time



[6] For our discussion below, variation over time is important. If always the same share of decisions were reached through consensus voting, omitting these votes would still understate the similarity of co-voting patterns but would not affect the comparability of affinity values over time.

[7] Hug (2012) shows that even in other than resolution-related votes there is considerable variation in the share of adoptions without votes.

3 Affinity measures and consensus voting

The problem generated by consenus votes 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 non-recorded votes. However, in the case of bodies of IOs, the lack of an explicit vote signals consensus among the delegates (see Lockwood Payton 2010, 2011). While a consensus decision is obviously not exactly the same thing as a unanimous endorsement of a proposal, an adoption without a vote suggests at least a broad acceptance of the decision among the delegates present. Yet if in one year 50 percent of the votes are omitted in the calculation of affinity because they were consensual, and in the next year 90 percent of the votes are omitted because they were consensual, observed changes in the values of those measures over time become largely meaningless.[8] Nevertheless, such affinity measures have become very popular in various subfields. For example, Gartzke (1998, see also Gartzke 2000, 2007) draws heavily on them when dealing with explanations of conflict. Alesina and Dollar (2000, see also Alesina & Weder 2002) 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 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 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 (see Stokman 1977 and Mokken & Stokman 1985 for similar suggestions in the context of UNGA voting).

Bailey, Strezhnev and Voeten (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 persistant disagreement might not appear in the following year, simply because the conflict has

[8] This is also implicitly acknowledged by the US State Department which has started in 1990 to assess voting coincidence not only for important votes (mostly on resolutions) but by including important consensus actions (i.e., adoptions without a vote) as well (See *Voting practices in the United Nations* 1990, US State Department, p. 220).

subsided and no resolutions addressed this conflict. They propose to overcome this problem by using an Item-Response Theory (IRT)-model, which alows to estimate 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 configurations in the ideal points can be considered as changes in preferences, and the distances among governments give indication of how close or far apart particular pairs of countries are.

However, this way to proceed is not without criticism, as the pertinence of the bridging observations is based on very strong assumptions. For instance, it assumes that the scales being estimated are actually the same from one year to the next and that the way in which they translate into votes for the bridging observations is actually the same. Jessee (2010) as well as Lewis, Jeffrey and Tausanovitch (2013) assess some recent studies from the American context of Congress employing a similar strategy and find that the necessary assumptions are almost never fulfilled. In addition, the IRT-models used in this context do not consider consensus votes, as the latter offer no information for estimating the parameters in an IRT-model.

4 Accounting for consensus voting in affinity measures

Having discussed the problem of consensus voting and its prevalence in the UNGA, we now turn to a more detailed discussion on why consensual votes generate biases in affinity measures. We assume that a resolution adopted without a vote had the implicit support of all members of the UNGA at the time of the vote.[9]Obviously, this is a strong assumption and errs in the direction of finding higher levels of similarity, but this overestimation is likely to be much smaller than the underestimation caused by ignoring the entire share of consensual decisions. In this section, we argue that the neglect of consensual votes in the calculation of vote agreement indices is justified neither on conceptual nor methodological grounds. We

[9] The affinity measures are not affected by the way consensual votes are coded, as long as they are coded in the same way for all member states. Assuming that a consensual 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 of course more substantive sense. Hovet (1960) includes in his analysis also non-recorded votes by relying on information obtained from UN embassy staff. It is unclear, however, whether this information also covers adoptions without vote and how reliable this information is.

also illustrate how the neglect of consensual votes leads to generally biased agreement values as well as problems regarding their comparability over time.

4.1 The effect of ignoring consensual 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).[10] 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 $i,j = 1, ..., k$ 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:

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

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 behaviour 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 (e.g., Lijphart 1963: 910; Gartzke 1998: 14-15, but see Voeten 2000: 193). Thus, weighting the 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 (i.e. yea or nay) and the middle category (i.e. abstention) seems justified. Figure 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, i.e. $w_{ij} = |i - j|$. This

[10] 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.

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 subjected to a conscious decision by the researcher.[11] 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:

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

The weights for the individual cells given our particular weighting function are shown in Figure 2. 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 $w_{max} = |3 - 1| = |1 - 3| = 2$.

[11] 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, i.e. $w_{ij} = (i - j)^2$. 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.

Figure 2: Calculation of proportion of disagreement for ordinal variables

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

Table 1 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 behaviour of states on resolutions adopted during that time period.[12] The table presents data for two years, with ten resolutions adopted in each of them, and information about the voting behaviour 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 behaviour of the five permanent UN Security Council members during the Cold War. During that period of time, the USA had diametrically opposed interests to the USSR, the UK and France were more closely aligned with the US, and China had more interests in common with the USSR.[13] The rows of the table with a grey background indicate resolutions adopted by consensus. Existing measures of vote agreement ignore these types of resolutions.

The arbitrariness of the neglect of consensual 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 behaviour of the two states that are members of the particular dyad. In our example, only the information provided in the USA and USSR columns of Table 1 are of relevance for calculating the dyadic, year-specific vote agreement value for those two countries (as highlighted by the heavy-bordered rectangle). As the voting behaviour of third parties is irrelevant for the calculation of the proportion of disagreement, no compelling

[12] 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.

[13] The extent to which the artificial data in Table 1 do indeed reflect the actual voting behaviour of those states during the Cold War is incidental to the argument we make here.

reason exists to exclude resolutions on which both the US and the USSR voted in favour, just because all other states voted in favour as well. Consider the first four resolutions of year 1. In all four cases, both the USA and the USSR voted in favour of the resolution. Yet when consensual decisions are excluded from the dataset, the voting behaviour on the first two resolutions is discarded. From a measurement point of view, given how the proportion of disagreement is defined, the voting behaviour on the first two resolutions provide exactly the same information for the calculation of the proportion of disagreement between the US and the USSR than the third and fourth resolution.

Table 1: 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

Notes: 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 ten 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 grey 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.

Ignoring resolutions adopted by consensus has non-trivial 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 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. Figure 3 illustrates these problems with our example data from Table 1. 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 behaviour in year 1 and the right column of contingency tables on the voting behaviour 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, i.e. in the way third states voted. In year 1, two out of ten decisions (i.e. 20 per cent) were taken by consensus. In contrast, in year 2, four out of ten decisions (i.e. 40 per cent) 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 well. Indeed, when consensual decisions are taken into account in its calculation, 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 consensual 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.

Figure 3: Consequences of excluding consensual decisions

Year 1

A. Consensual decisions included

		USA			
		1	2	3	Total
USSR	1	0 (.00) 0	1 (.10) 1	2 (.20) 2	3 (.30)
	2	0 (.00) 1	1 (.10) 0	1 (.10) 1	2 (.20)
	3	1 (.10) 2	0 (.00) 1	4 (.40) 0	5 (.50)
	Total	1 (.10)	2 (.20)	7 (.70)	10 (1)

$D_o^1 = \frac{0.80}{2} = 0.4$

B. Consensual decisions excluded

		USA			
		1	2	3	Total
USSR	1	0 (.00) 0	1 (.125) 1	2 (.25) 2	3 (.375)
	2	0 (.00) 1	1 (.125) 0	1 (.125) 1	2 (.25)
	3	1 (.125) 2	0 (.00) 1	2 (.25) 0	3 (.375)
	Total	1 (.125)	2 (.25)	5 (.625)	8 (1)

$D_o^1 = \frac{1}{2} = 0.5$

Year 2

A. Consensual decisions included

		USA			
		1	2	3	Total
USSR	1	0 (.00) 0	1 (.10) 1	2 (.20) 2	3 (.30)
	2	0 (.00) 1	1 (.10) 0	1 (.10) 1	2 (.20)
	3	1 (.10) 2	0 (0) 1	4 (.40) 0	5 (.50)
	Total	1 (.10)	2 (.20)	7 (.70)	10 (1)

$D_o^1 = \frac{0.80}{2} = 0.4$

B. Consensual decisions excluded

		USA			
		1	2	3	Total
USSR	1	0 (.00) 0	1 (.17) 1	2 (.33) 2	3 (.50)
	2	0 (.00) 1	1 (.17) 0	1 (.17) 1	2 (.33)
	3	1 (.17) 2	0 (.00) 1	0 (.00) 0	1 (.17)
	Total	1 (.17)	2 (.33)	3 (.50)	6 (1)

$D_o^1 = \frac{1.33}{2} = 0.67$

Notes: The tables are based on the artificial data presented in Table 1. 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 excluded from 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: $D_o^1 = (0 * 0 + 1 * 0.1 + 2 * 0.2 + 1 * 0 + 0 * 0.1 + 1 * 0.1 + 2 * 0.1 + 1 * 0 + 0 * 0.4 = 0.1 + 2 * 0.2 + 0.1 + 2 * 0.1)/2 = (0.1 + 0.4 + 0.1 + 0.2)/2 = 0.8/2$

In this particular example, the proportion of disagreement is 0.40 in both years when consensual decisions are included.[14] 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: proportion 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 towards more disagreement.

4.2 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 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. In general, any chance-corrected agreement index A takes the following form:

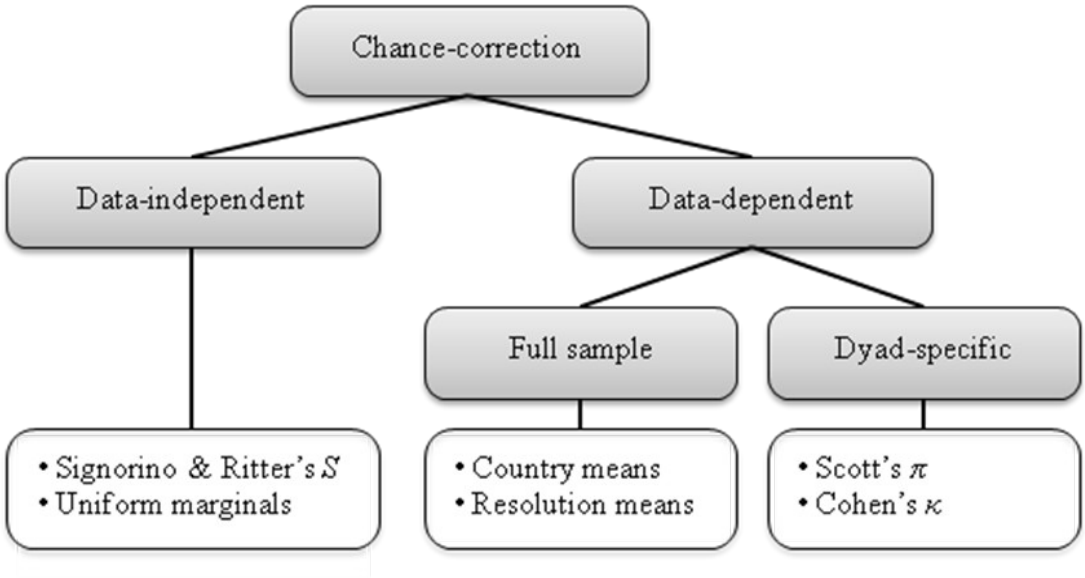
$$A = 1 - \frac{D_o}{D_e}$$

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.

[14] See the notes to Figure 3 for a detailed example of how the proportion of disagreement is calculated from the information in the contingency tables.

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 4 shows the resulting classification tree.

Figure 4: Classification of chance-correction approaches



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 $l = 1, ..., r$. Thus, for each resolution, *S* first calculates the 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:

$$S = 1 - 2 * \sum_{l=1}^{r} \frac{|y_l - x_l|}{d_{max}} = 1 - 2 * \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{w_{ij} f_{ij}}{w_{max} f_{ij}} = 1 - 2 * \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} f_{ij}}{w_{max} f_{..}}$$

$$= 1 - 2 * \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}}{w_{max}} = 1 - 2 * D_o = 1 - \frac{D_o}{0.5}$$

The reformulation makes it clear that *S* is simply a linear function of the proportion of disagreement D_o. 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 ½. 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:

$$D_e = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{i\cdot} m_{\cdot j}}{w_{max}}$$

Different indices vary only in the assumptions they make about the marginals $m_{i\cdot}$ and $m_{\cdot j}$ 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. Table 2 summarizes these assumptions for the agreement indices discussed in this section.

Table 2: Assumptions about marginal distributions for chance-correction

Index	Assumptions about marginal distributions
Signorino & Ritter's S	$m_{1\cdot} = m_{\cdot 1} = 0.5$ $m_{2\cdot} = m_{\cdot 2} = 0.0$ $m_{3\cdot} = m_{\cdot 3} = 0.5$
Uniform marginals	$m_{i\cdot} = m_{\cdot j} = \frac{1}{3}$ for $i,j = 1, 2, 3$
Resolution average marginals	$m_{i\cdot} = m_{\cdot j} = \frac{\sum_{l=1}^{r} p_{li\cdot}}{r}$ for $i,j = 1, 2, 3$ and $l = 1, ..., r$, where r stands for the number of resolutions
Country average marginals	$m_{i\cdot} = m_{\cdot j} = \frac{\sum_{g=1}^{n} p_{gi\cdot}}{n}$ for $i,j = 1, 2, 3$ and $g = 1, ..., n$, where n stands for the number of member states
Scott's π	$m_{i\cdot} = m_{\cdot j} = \frac{p_{i\cdot} + p_{\cdot j}}{2}$ for $i,j = 1, 2, 3$
Cohen's κ	$m_{i\cdot} = p_{i\cdot}$ for $i = 1, 2, 3$ $m_{\cdot j} = p_{\cdot j}$ for $j = 1, 2, 3$

Figure 5 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. 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 (i.e. yea or nay) and a zero propensity to fall into the intermediate category (i.e. abstain). However, if we assume that both member states have the same propensities to vote in a certain way, i.e. 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 Figure 3.

The assumptions about the form of the marginal distributions used to calculate the chance correction of *S* are hard to justify on substantive grounds.[15] Assuming that states have a 50 per cent probability of voting 'yea' or 'nay' and a zero per cent probability of abstaining

[15] Mokken and Stokman (1985: 187-8) argue that this chance correction is useful for measuring the cohesion of a decision-making body as a whole.

contradicts both common sense and available empirical information.[16] 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 'yeah', 'nay' or abstain (e.g. Lijphart 1963: 906-8, Mokken and Stokman 1985: 186-7). Panel C of Figure 5 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 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 (i.e. identical for both states).[17] With the exception of Cohen's κ, all of the indices discussed here make this assumption.

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 behaviour. Another way of specifying the values for the marginal distributions is to estimate them from the information in the sample. Mokken and Stokman (1985: 187) propose to estimate the marginals by computing for each resolution the proportion of states voting in favour, 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 Figure 5). 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 toy example of Figure 5, 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 might not have 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

[16] 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.

[17] 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.

non-uniform 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 consensual votes are part of the sample.

Figure 5: Calculation of indices based on different assumptions about marginals

A. Observed disagreement

		USA 1	USA 2	USA 3	
USSR	1	0.000	0.101	0.202	0.30
	2	0.001	0.100	0.101	0.20
	3	0.102	0.001	0.400	0.50
		0.10	0.20	0.70	1

$D_o = (4 * 0.10 + 2 * 0.20)/2 = 0.40$

B. Signorino and Ritter's *S*

		USA 1	USA 2	USA 3	
USSR	1	0.250	0.001	0.252	0.50
	2	0.001	0.000	0.001	0.00
	3	0.252	0.001	0.250	0.50
		0.50	0.00	0.50	1

$D_e^S = (4 * 0.25)/2 = 0.50$

$A_e^S = 1 - \frac{D_o}{D_e^S} = 1 - \frac{0.40}{0.50} = 0.2$

C. Uniform marginals

		USA 1	USA 2	USA 3	
USSR	1	0.110	0.111	0.112	0.33
	2	0.111	0.110	0.111	0.33
	3	0.112	0.111	0.110	0.33
		0.33	0.33	0.33	1

$D_e^u = (8 * 0.11)/2 = 0.44$

$A_e^u = 1 - \frac{D_o}{D_e^u} = 1 - \frac{0.40}{0.44} = 0.10$

D. Country/resolution average marginals

		USA 1	USA 2	USA 3	
USSR	1	0.030	0.051	0.102	0.18
	2	0.051	0.080	0.151	0.28
	3	0.102	0.151	0.290	0.54
		0.18	0.28	0.54	1

$D_e^v = \left(\begin{matrix}2 * 0.05 + 4 * 0.10 \\ +2 * 0.15\end{matrix}\right)/2 = 0.40$

$A_e^\kappa = 1 - \frac{D_o}{D_e^v} = 1 - \frac{0.40}{0.40} = 0.00$

D. Scott's τ

		USA 1	USA 2	USA 3	
USSR	1	0.040	0.041	0.122	0.20
	2	0.041	0.040	0.121	0.20
	3	0.122	0.121	0.360	0.60
		0.20	0.20	0.60	1

$D_e^\pi = (2 * 0.04 + 6 * 0.12)/2 = 0.40$

$A_e^\pi = 1 - \frac{D_o}{D_e^\pi} = 1 - \frac{0.40}{0.40} = 0.00$

E. Cohen's κ

		USA 1	USA 2	USA 3	
USSR	1	0.030	0.061	0.212	0.30
	2	0.021	0.040	0.141	0.20
	3	0.052	0.101	0.350	0.50
		0.10	0.20	0.70	1

$D_e^\kappa = \left(\begin{matrix}0.06 + 2 * 0.21 + 0.14 + \\ 0.02 + 2 * 0.05 + 0.10\end{matrix}\right)/2 = 0.84$

$A_e^\kappa = 1 - \frac{D_o}{D_e^\kappa} = 1 - \frac{0.40}{0.42} = 0.05$

However, when it comes to agenda effects, both the voting behaviour of particular dyads and individual countries within dyads might be unduly affected as well. Scott's (1955) π and Cohen's κ (1968) address these issues. The country average marginals approach is basically an extension of the chance-correction approach used in the calculation of Scott's π. 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 π only averages the vote propensities of the two states that form part of the particular dyad. In this respect, Scott's π is more flexible and able to not only adjust for factors that affect the voting behaviour of all states in the sample equally (e.g. consensual votes), but also for factors that affect only the voting behaviour of the particular dyad members in the same way. Yet Scott's π still assumes that both dyad members have the same baseline propensities to vote in a certain way, although good reasons exist to expect that certain factors have divergent effects on the voting behaviour of dyad members. For some dyads, a certain agenda might lead to dyad members voting the same way more often, for other dyads, the same agenda might lead to their members voting in opposite ways more often. Cohen's κ goes a step further than Scott's π and allows each dyad member to have its own independent marginal distribution for the calculation of the proportion of expected disagreement. This measure directly uses the marginal distributions of the observed contingency table to estimate the expected marginal distributions. Given that Cohen's κ is most versatile in adjusting for both the inclusion of consensual votes and the potentially divergent effects on voting behaviour resulting from changes in the agenda, the following replication studies focus on the performance of this chance-corrected agreement index.

5 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 κ and Scott's π 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 κ or π 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 liteature on foreign aid. In a path breaking 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 US (see also Alesina & Weder 2002). In what follows we carry out replications of two models of Alesina and Dollar's (2000) study, namely explanations of total bilateral aid and US bilateral aid as given in five year periods to recipient countries.[18] 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.[19]

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 Alker and Russett's (1965) data (see also Strezhnev & Voeten 2012). Unfortunately, combining data from different sources has led to a situation in which the inclusion criteria vary across time periods (e.g., votes on amendments etc. are included until the 1970s, but figure no longer in the data for more recent periods; for a related discussion, see Rai 1982). For this reason we rely on Hug's (2012) data (for a publication using this data, see Hug & Wegmann 2013), which comprises, based on a common source, all votes on resolutions as well as information on all resolutions debated in the UNGA (for a similar effort, see Skougarevskiy 2012). As we have both information on all votes related to resolutions as well as information on resolutions adopted without a vote, we proceed as follow:

- First, we generate for each year a dataset that only comprises the member state voting records on resolutions.
- Second, we generate an imputed dataset where for all states that were members of the UN at the time of the vote, we assume that they voted in favour of all resolutions adopted without a vote.[20]

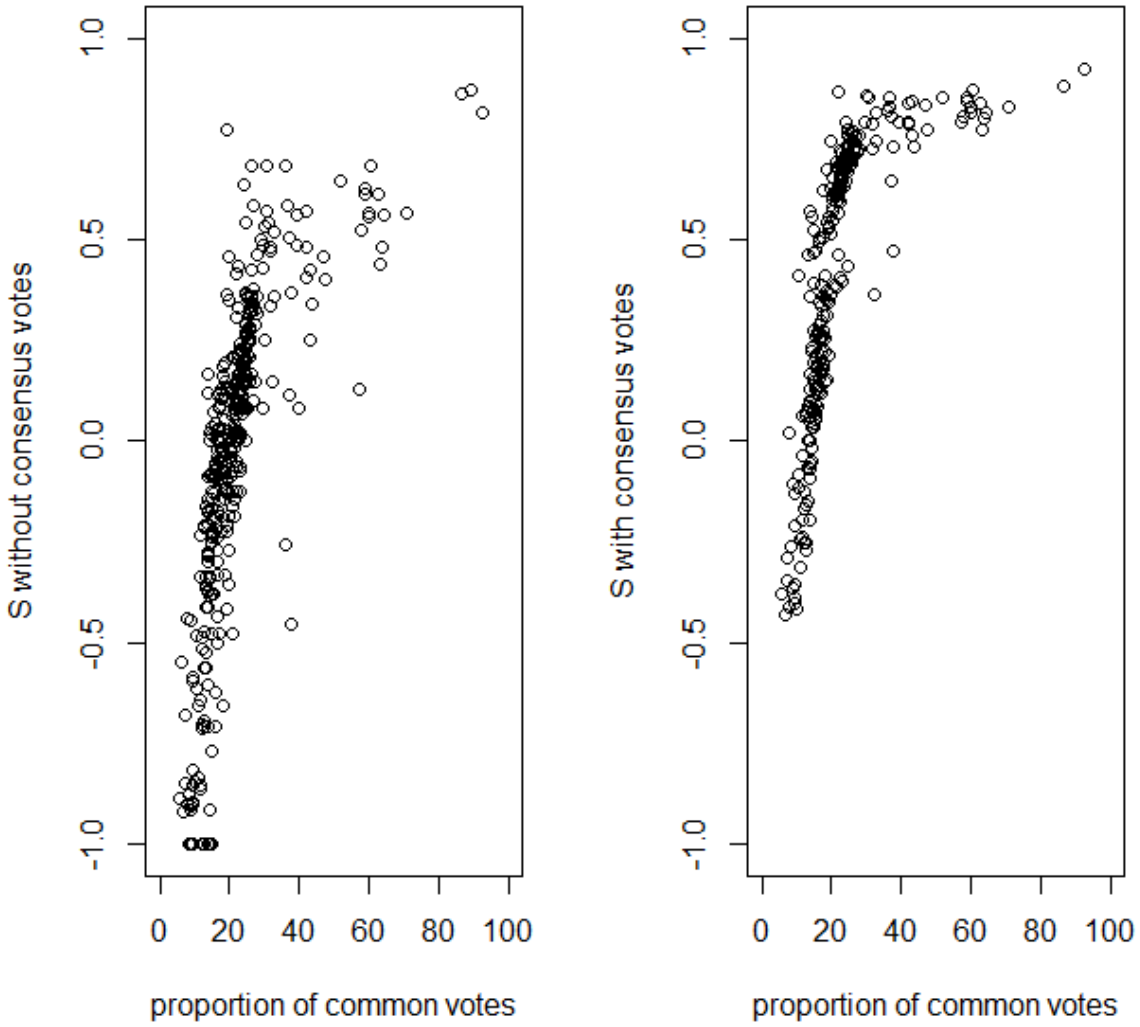
As Alesina and Dollar's (2000) study uses five-year averages as the unit of analysis for aid allocation and all other variables, we also aggregated our yearly similarity measures based on

[18] We obtained the replication data from http://aiddata.org/content/index/Research/replication-datasets website.
[19] As with almost all other measures, the authors offer almost no explanation of how this measure was constructed. For instance, we do not know whether abstentions were counted, whether proportions were calculated over the entire five-year period or for individual years and then somehow aggregated over the five-year period.
[20] Again, it is important to note that we make an assumption, namely that adoptions without a vote signal unanimous support of the resolution adopted.

our imputed UNGA voting data by calculating five year averages. We then merged our data with Alesina and Dollar's (2000) replication dataset. As a first 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). In Figure 6 and 7, we depict this relationship by using either Signorino and Ritter's (1999) *S* or Cohen's κ, while varying whether or not we include consensus votes.

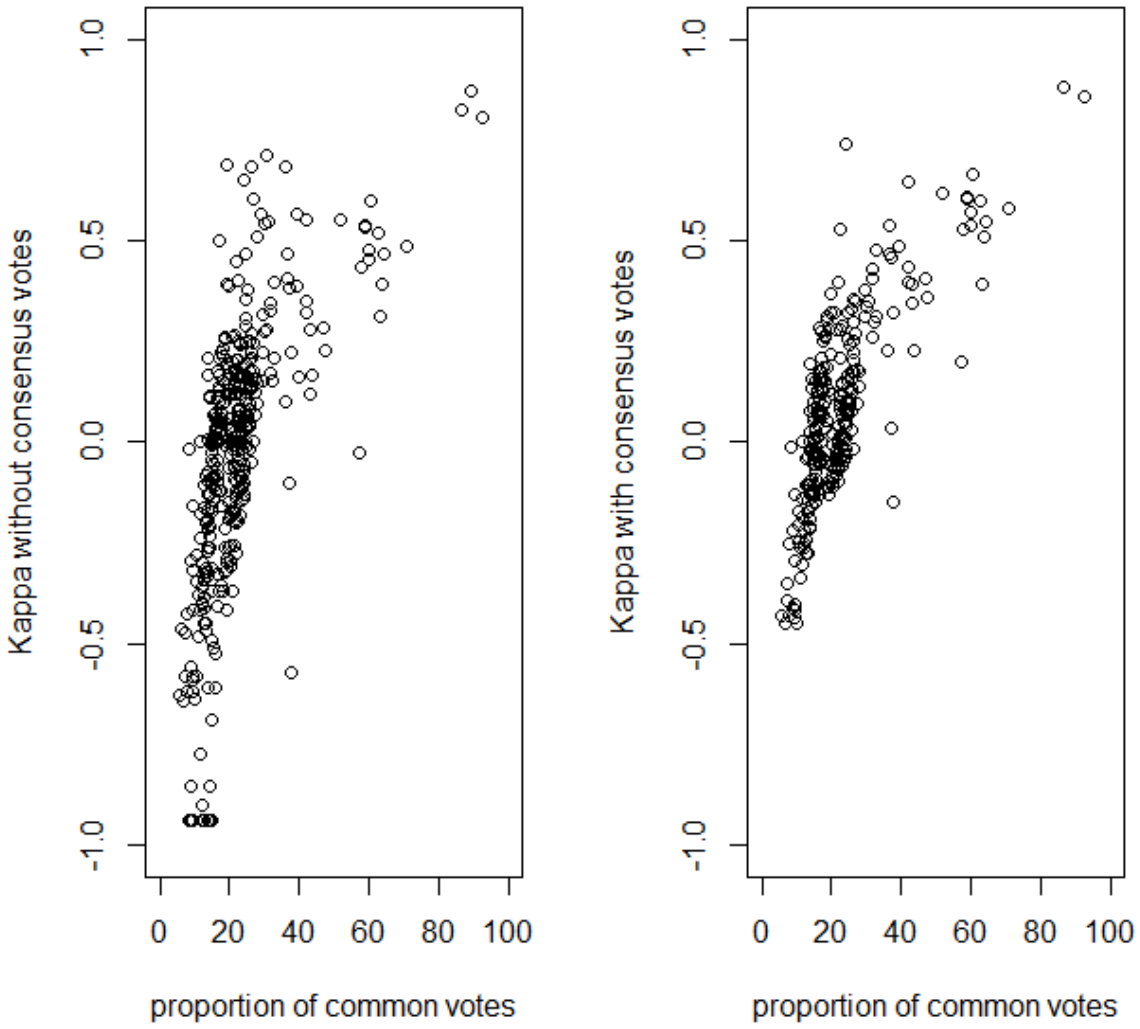
In Figure 6, where we compare *S* to the proportion of common votes, we find that in the first panel, i.e. without consensus votes, the two measures are closely related. Given that *S* is a linear transformation of the proportion of common votes, this is not surprising. Indeed, any deviation from a perfect relationship between the two variables must be due to differences in the underlying data. When taking consensus votes into account (second panel in Figure 6), we find much higher similarity values, but also a much weaker relationship, with considerable variation around the average trend.

Figure 6: Affinity measures (5 year averages) for the United States (I)



In Figure 7, where we rely on Cohen's κ, already the first panel (omitting consensus votes) shows a rather weak relationship. Again, once we include consensus votes the average value of κ increases and the relationship with Alesina and Dollar's (2000) proportion of common votes becomes considerably blurred. Hence, it is very likely that this proportion of common votes, by not considering consensus votes, is actually measuring something that is hardly meaningful.

Figure 7: Affinity measures (5 year averages) for the United States (II)



We assess this by re-estimating two of Alesina and Dollar's (2000) models, namely one explaining the total bilateral aid obtained in five-year periods by aid recipients (Table 3) and the other focusing only on US bilateral aid (Table 4). While Alesina and Dollar (2000) make their data available, there are very few indications on how this data was used to produce the results reported in their paper. Thus, in both tables 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 Japan, respectively) with *S* and κ. In the first two models , the

affinity measures are based only on recorded votes, in the last two models we also include information on consensual votes in the calculation of *S* and κ.

Considering Table 3 first, we find that, as in Alesina and Dollar's (2000) analysis, trade-openness considerably increases aid allocations. Similarly, having been a colony for a longer time or being either Egypt or Israel increases aid significantly independent of the similarity measure used (i.e., in all models in Table 3). When it comes to the political variables, however, results prove to be less stable. While, consistent with Alesina and Dollar (2000), we find a positive effect of political liberties on aid, the effects of voting similarities with the US and Japan are far from robust. While we can reproduce the statistically significant effect of voting similarity with Japan on bilateral aid and, conversely, the absence of an effect for voting with the US, these results are very sensitive to the measure and data used. For instance if κ is used instead of *S*, independent of whether consensus votes are considered or not, closeness to Japan appears not to matter. On the other hand if we consider *S* as similarity measure, closeness to the US almost appears to have a statistically significant effect on aid allocations.

In Table 4 we report the results of our second replication that focuses on explaining US bilateral aid.[21]In this replication we are unable to reproduce the positive effect of GDP per capita reported by Alesina and Dollar (2000).[22] For the remainder of the variables we are able to approximate the results, except that no former colony of the US has non-missing data on all variables, which is the reason for which this variable drops from our replication.

We are able to only partly replicate the positive effect of voting with the US on obtaining aid from this country. When we consider *S* and κ as similarity measures while ignoring consensus votes, the effect is positive and statistically significant for both variables. However, when considering consensus votes in the calculation of those similarity measures as well, only the effect of κ remains statistically significant.

[21] Despite having a larger number of cases in our replication of Alesina and Dollar's (2000) model on US bilateral aid than in their original published study, we find in our sample, based on Alesina and Dollar's (2000) data, no cases that were US colonies. Consequently, this variable drops from our analysis.
[22] Given the robustness of the negative effect of this variable in the remainder of the models in Table 4, we can only suspect a typo in the Alesina and Dollar (2000) article.

Table 3: Replication of Alesina and Dollar (2000), total bilateral aid I

	similarity measure					
	without consensus votes				with consensus votes	
	proportion of agreement		chance corrected			
			S	K	S	K
	b	b	b	b	b	b
	(t)	(se)	(se)	(se)	(se)	(se)
Log GDP per capita	6.563 *	7.070 *	7.154 *	6.635 *	7.095 *	6.568 *
	(4.77)	(1.106)	(1.122)	(1.174)	(1.168)	(1.226)
Log GDP per capita^2	-0.491 *	-0.598 *	-0.603 *	-0.570 *	-0.599 *	-0.566 *
	(5.32)	(0.074)	(0.075)	(0.078)	(0.078)	(0.082)
Log population	1.568	0.160	0.225	0.324	0.144	0.238
	(1,91)	(0.478)	(0.478)	(0.499)	(0.487)	(0.509)
Log population^2	-0.035	-0.023	-0.024	-0.027	-0.022	-0.024
	(1.36)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)
Economic openness	0.383 *	0.341 *	0.391 *	0.512 *	0.409 *	0.535 *
	(2.57)	(0.158)	(0.161)	(0.167)	(0.166)	(0.172)
Democracy	0.142 *	0.134 *	0.116 *	0.130 *	0.108 *	0.123 *
	(3.23)	(0.035)	(0.036)	(0.038)	(0.038)	(0.040)
Friend of USA (UNGA voting)	-0.006	-0.006	-0.620	2.492	-0.673	2.408
	(0.30)	(0.009)	(0.661)	(1.286)	(0.672)	(1.316)
Friend of Japan (UGA voting)	0.153 *	0.086 *	5.400 *	0.621	5.516 *	0.705
	(4.0)	(0.015)	(0.853)	(0.780)	(0.866)	(0.798)
Log years as colony	0.291 *	0.217 *	0.219 *	0.236 *	0.215 *	0.233 *
	(4.64)	(0.041)	(0.041)	(0.043)	(0.042)	(0.044)
Egypt	1.545 *	1.594 *	1.648 *	1.686 *	1.650 *	1.690 *
	(10.53)	(0.504)	(0.501)	(0.523)	(0.506)	(0.530)
Israel	6.473 *	6.077 *	5.764 *	3.528 *	5.853 *	3.601 *
	(3.03)	(0.772)	(0.723)	(0.715)	(0.734)	-0.73
Percent muslims	-0.001	0.006 *	0.006 *	0.007 *	0.006 *	0.008 *
	(0.42)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Percent catholics	0.001	0.008 *	0.007 *	0.007 *	0.007 *	0.008 *
	(0.30)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Percent other religions (Hindu)	-0.009 *	0.003	0.004	0.004	0.004	0.005
	(2.94)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
1970-1974		-25.230 *	-22.819 *	-18.575 *	-22.052 *	-17.679 *
		(5.802)	(5.886)	(6.131)	(6.030)	(6.299)
1975-1979		-26.002 *	-22.389 *	-18.338 *	-21.595 *	-17.422 *
		(5.827)	(5.885)	(6.131)	(6.028)	(6.299)
1980-1984		-24.717 *	-21.218 *	-18.155 *	-20.425 *	-17.244 *
		(5.821)	(5.886)	(6.132)	(6.029)	(6.300)
1985-1989		-24.893 *	-21.068 *	-17.680 *	-20.284 *	-16.766 *
		(5.843)	(5.892)	(6.130)	(6.035)	(6.297)
1990-1994		-24.991 *	-21.294 *	-17.604 *	-20.524 *	-16.701 *
		(5.845)	(5.900)	(6.140)	(6.043)	(6.307)
N	397	402	386	386	372	372
R^2	0.63	0.806	0.814	0.797	0.805	0.787
Adj. R^2	0.978	0.978	0.971	1.014	0.981	1.027

Standard errors in parentheses
* indicates significance at $p<0.05$

Table 4: Replication of Alesina and Dollar (2000), bilateral aid by US I

	similarity measure					
	without consensus votes				with consensus votes	
	proportion of agreement		chance corrected			
			S	K	S	K
	b	b	b	b	b	b
	(t)	(se)	(se)	(se)	(se)	(se)
Log GDP per capita	1.84	-1.647 *	-1.718 *	-1.692 *	-1.688 *	-1.664 *
		(0.151)	(0.151)	(0.151)	(0.151)	(0.151)
Economic openness	1.30 *	0.817 *	0.808 *	0.799 *	0.805 *	0.795 *
	(4.02)	(0.281)	(0.287)	(0.286)	(0.285)	(0.284)
democracy	0.57 *	0.387 *	0.445 *	0.426 *	0.475 *	0.459 *
	(8.07)	(0.063)	(0.066)	(0.067)	(0.066)	(0.066)
Friend of USA (UNGA voting)	0.06 *	0.043 *	2.105 *	3.883 *	1.787	3.349 *
	(3.60)	(0.014)	(0.997)	(1.386)	(0.990)	(1.386)
Log years as colony of US	0.39					
	(1.69)					
Log years as colony not of US	-0.007	-0.007	-0.009 *	-0.009 *	-0.010 *	-0.010 *
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Egypt	4.502 *	4.502 *	4.444 *	4.425 *	4.402 *	4.383 *
	(0.882)	(0.882)	(0.877)	(0.872)	(0.863)	(0.859)
Israel	4.692 *	4.692 *	5.403 *	5.112 *	5.541 *	5.268 *
	(1.099)	(1.099)	(1.107)	(1.054)	(1.095)	(1.046)
Percent muslims	0.022 *	0.022 *	0.023 *	0.023 *	0.024 *	0.024 *
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Percent catholics	0.018 *	0.018 *	0.018 *	0.018 *	0.020 *	0.019 *
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Percent other religions (Hindu)	0.011	0.011	0.009	0.011	0.009	0.010
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
1970-1974		9.330 *	11.627 *	11.791 *	11.339 *	11.470 *
		(1.311)	(1.088)	(1.065)	(1.080)	(1.059)
1975-1979		9.238 *	11.874 *	11.861 *	11.522 *	11.502 *
		(1.323)	(1.086)	(1.078)	(1.081)	(1.075)
1980-1984		10.928 *	13.002 *	12.188 *	12.513 *	11.816 *
		(1.138)	(1.132)	(1.084)	(1.128)	(1.082)
1985-1989		11.084 *	13.104 *	12.067 *	12.637 *	11.750 *
		(1.108)	(1.165)	(1.077)	(1.159)	(1.074)
1990-1994		10.583 *	12.785 *	12.007 *	12.416 *	11.749 *
		(1.138)	(1.136)	(1.079)	(1.132)	(1.077)
N	364	364	349	349	338	338
R^2	0.5	0.713	0.718	0.721	0.732	0.734
Adj. R^2		1.702	1.690	1.681	1.662	1.655

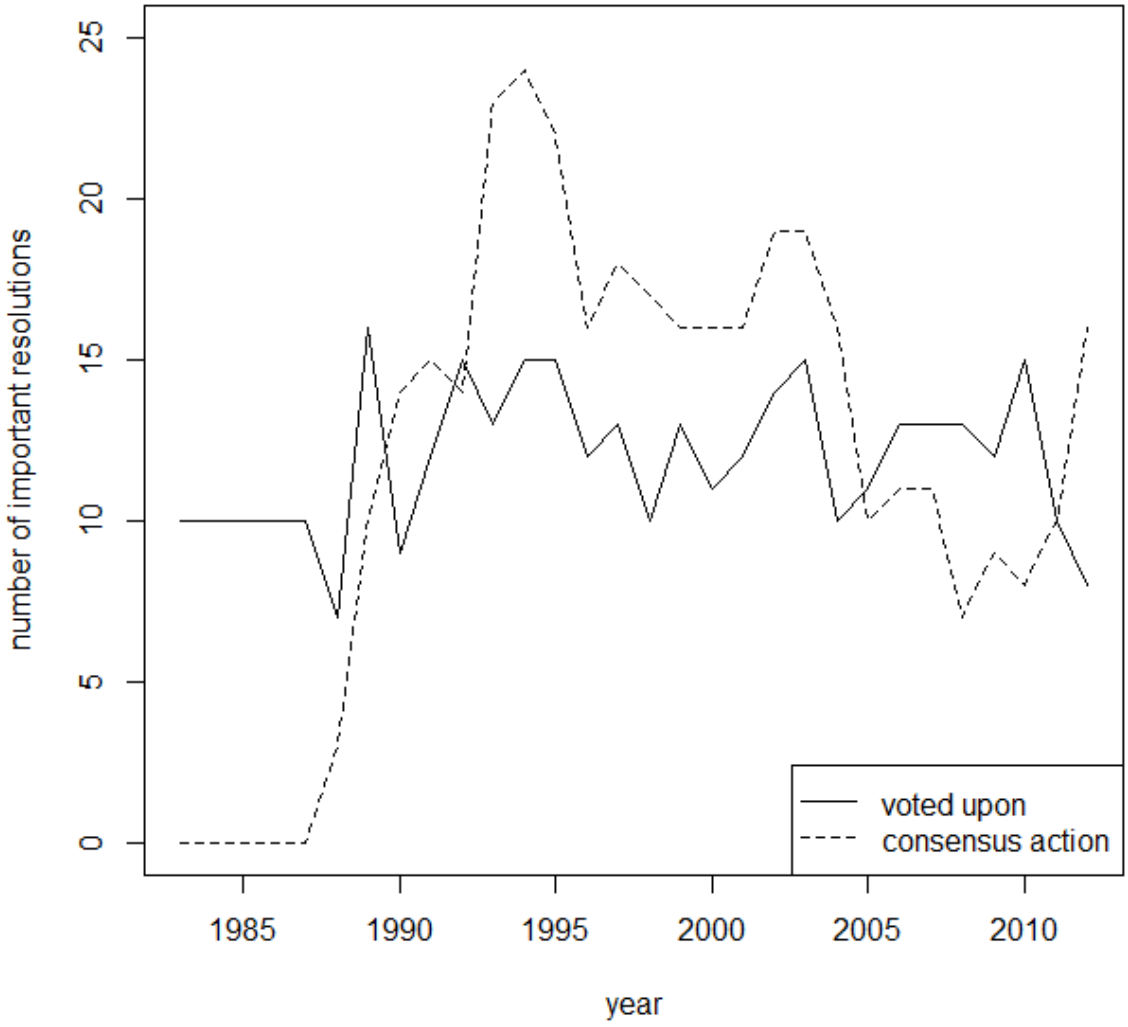
Standard errors in parentheses
* indicates significance at $p<0.05$

Consequently, the political and strategic explanations advocated by Alesina and Dollar (2000) appear to be on rather weak empirical footings. We find only mixed evidence that being close to Japan in UNGA voting significantly increases the overall reception of bilateral aid. Also, the effect of voting similarity with the US in explaining US bilateral aid seems far from robust. When using both consensus votes and a chance correction, only voting similarity as measured by κ appears to significantly affect US bilateral aid.

6 Discussion

A valid criticism of our apoproach is that consensus votes might simply be secondary or even unimportant decisions. One way to assess this is to consider a commonly used source to restrict votes to the important ones. 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 ten recorded votes (most of these are final passage votes of resolutions) that it deemed to be "key," respectively important (for a list see the appendix in Thacker 1999). In 1988 the state department designated, however, three decisions reached without a vote as equally important. Starting from then on the reports by the state department list both important votes and important "consensus actions." In Figure 8 we depict the number of important votes and consensus actions from 1983 to 2012. As this figure clearly shows for large periods of time the US state department considers more consenus actions (i.e., matters adopted without a vote) as important than matters adopted in a recorded 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.

Figure 8: Proportion of important resolutions in the UNGA over time



Obviously, when considering Figure 8, however, one has also to take into account the number of recorded votes and adoption without votes. Nevertheless, based on this figure, it is hard to argue that decisions reached without a vote are systematically unimportant. However, as adoptions without votes are generally more prevalent than recorded votes, Figure 8 suggests that the relative share of important resolutions (as assessed by the US state department) is higher among recorded votes than among consensus votes. To assess whether important decisions among consensus votes can be ignored, we carried out an additional set of replication analyses of Alesina and Dollar's (2000) study. More precisely, we reestimated the models reported in Table 3 and 4 with similarity measures calculated only on the basis of

important resolutions. As the US State Department only started in 1983 to report on these important resolutions, we limit our analysis to the 5 year periods starting in 1985.[23]

The first two columns in Table 5 are identical to those in Table 3 and present first the results reported in Alesina and Dollar (2000) and second our replication. The third model is identical to the one used for the results of the second model, but the period covered for the similarity measures only starts in 1985. For this same period we then estimated exactly the same models as those reported in Table 3. As the data used for these models is sparser (only two 5-year periods), it is not surprising that our estimates are on average less precise. In terms of substance, Table 5 offers largely the same results, with one important exception. As in Table 3, the effect of voting with Japan only appears for Signorino and Ritter's (1999) *S* but not for κ. However, we suddendly find a negative and significant effect for US voting similarlity based Signorino and Ritter's (1999) *S*, provided we use only important votes and no consensus actions. Yet this effect disappears if we use κ as similarity measure or if we also consider consensus actions.

[23] As Figure 8 shows, in the first years the US State Department restricted its reporting of important resolutions to those adopted through a vote. Thus, starting our analysis in 1985 instead of 1990 biases our results against finding a difference between similarity measures based only on contested votes and those also taking into account consensus actions.

Table 5: Replication of Alesina and Dollar (2000), total bilateral aid II

	similarity measure						
	without consensus votes					with consensus votes	
	proportion of agreement			chance corrected			
				S	K	S	K
	1965-1995		1985-1995				
	b	b	b	b	b	b	b
	(t)	(se)	(se)	(se)	(se)	(se)	(se)
Log GDP per capita	6.563 *	7.070 *	6.969 *	7.334 *	6.890 *	7.411 *	6.702 *
	(1.106)	(1.106)	(1.826)	(1.813)	(1.899)	(1.806)	(1.871)
Log GDP per capita^2	-0.491 *	-0.598 *	-0.584 *	-0.603 *	-0.576 *	-0.607 *	-0.562 *
	(0.074)	(0.074)	(0.121)	(0.120)	(0.126)	(0.119)	(0.124)
Log population	1.568 *	0.160	0.033	-0.148	0.262	-0.081	0.275
	(0.478)	(0.478)	(0.696)	(0.687)	(0.710)	(0.679)	(0.690)
Log population^2	-0.035 *	-0.023	-0.019	-0.013	-0.025	-0.015	-0.025
	(0.015)	(0.015)	(0.021)	(0.021)	(0.021)	(0.020)	(0.021)
Economic openness	0.383 *	0.341 *	0.312	0.227	0.270	0.202	0.230
	(0.158)	(0.158)	(0.224)	(0.217)	(0.225)	(0.218)	(0.224)
Democracy	0.142 *	0.134 *	0.077	0.016	0.053	0.013	0.040
	(0.035)	(0.035)	(0.064)	(0.062)	(0.063)	(0.062)	(0.063)
Friend of USA (UNGA voting)	-0.006	-0.006	-0.005	-4.413 *	1.188	-2.348	1.535
	(0.009)	(0.009)	(0.025)	(1.751)	(1.764)	(1.223)	(1.067)
Friend of Japan (UNGA voting)	0.153 *	0.086 *	0.007	6.075 *	-0.335	3.591 *	-0.547
	(0.015)	(0.015)	(0.046)	(1.942)	(1.376)	(1.342)	(0.852)
Log years as colony	0.291 *	0.217 *	0.160 *	0.135 *	0.169 *	0.141 *	0.166 *
	(0.041)	(0.041)	(0.066)	(0.063)	(0.065)	(0.063)	(0.065)
Egypt	1.545 *	1.594 *	1.725 *	1.495 *	1.617 *	1.434 *	1.543 *
	(0.504)	(0.504)	(0.698)	(0.671)	(0.693)	(0.672)	(0.688)
Israel	6.473 *	6.077 *	5.427 *	6.575 *	4.095 *	6.425 *	3.796 *
	(0.772)	(0.772)	(2.609)	(1.017)	(1.162)	(1.126)	(0.966)
Percent muslims	-0.001	0.006 *	0.006 *	0.005	0.006	0.006 *	0.005
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Percent catholics	0.001	0.008 *	0.007	0.007 *	0.006	0.007	0.006
	(0.002)	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)
Percent other religions (Hindu)	-0.009 *	0.003	0.002	0.003	0.003	0.004	0.003
	(0.004)	(0.004)	(0.007)	(0.006)	(0.007)	(0.006)	(0.007)
1970-1974		-25.230 *					
		(5.802)					
1975-1979		-26.002 *					
		(5.827)					
1980-1984		-24.717 *					
		(5.821)					
1985-1989		-24.893 *	-17.017	-18.400 *	-18.182	-18.622 *	-17.653
		(5.843)	(9.856)	(9.077)	(9.417)	(9.068)	(9.334)
1990-1994		-24.991 *	-16.893	-18.061 *	-18.104	-18.567 *	-17.530
		(5.845)	(9.833)	(9.107)	(9.435)	(9.087)	(9.350)
N	402	402	155	151	151	151	151
R^2	0.806	0.806	0.818	0.833	0.821	0.833	0.824
Adj. R^2	0.796	0.796	0.797	0.813	0.799	0.813	0.803

Standard errors in parentheses
* indicates significance at $p<0.05$

Table 6 reports replications of the analyses depicted in Table 4, but this time basing the calculation of similarity measures on important resolutions only. Again, we find largely identical results for most of the variables, but also less precise estimates given the shorter time period covered. As for the total bilateral we find, however, also for bilateral aid by the US important differences resulting from different similarity measures. If we consider, as Alesina and Dollar (2000) do, the proportion of votes in agreement between the recipient country and the US, the effect is positive and statistically significant independent of the period covered. If we focus our analyses on important votes then the effects of *S* and κ are reduced and lose their statistical significance. If we also use, however, the consensus actions to calculate these two similarity measures, the two effects almost reach conventional levels of statistical significance. More precisely, the effects are not significant at the 0.05 level but cross the 0.10 level.

Thus, both of these two additional sets of replications underline our point that ignoring consensus actions, even when focussing only on matters before the UNGA deemed important by the US State Department, is perilous. Many of the results reported in Alesina and Dollar (2000) depend crucially on the omission of consensus actions to measure voting similarity and in addition to the use of rather problematic measures of voting similarities (see our discussion above as well as Signorino and Ritter 1999, Häge 2011, and Bailey, Strezhnev & Voeten 2013).

Table 6: Replication of Alesina and Dollar (2000), bilateral aid by US I

	similarity measure						
	without consensus votes					with consensus votes	
	proportion of agreement		chance corrected				
				S	K	S	K
	1965-1995		1985-1995				
	b	b	b	b	b	b	b
	(t)	(se)	(se)	(se)	(se)	(se)	(se)
Log GDP per capita	1.840 *	-2.306 *	-2.400 *	-2.423 *	-2.449 *	-2.395 *	-2.415 *
	(0.203)	(0.203)	(0.314)	(0.326)	(0.325)	(0.326)	(0.325)
Economic openness	1.300 *	1.246 *	0.855	0.936	0.977	0.903	0.915
	(0.378)	(0.378)	(0.561)	(0.580)	(0.580)	(0.579)	(0.579)
democracy	0.570 *	0.508 *	0.661 *	0.693 *	0.706 *	0.659 *	0.669 *
	(0.085)	(0.085)	(0.159)	(0.166)	(0.166)	(0.169)	(0.168)
Friend of USA (UNGA voting)	0.060 *	0.066 *	0.094 *	1.711	1.367	1.525	1.719
	(0.018)	(0.018)	(0.038)	(1.204)	(1.138)	(0.866)	(1.029)
Log years as colony not of US	0.080 *	-0.002	-0.000	-0.002	-0.001	-0.002	-0.001
	(0.005)	(0.005)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Egypt	40.090 *	5.071 *	6.024 *	5.834 *	5.915 *	5.754 *	5.798 *
	(1.188)	(1.188)	(1.705)	(1.741)	(1.742)	(1.735)	(1.735)
Israel	5.040 *	4.815 *	0.740	6.309 *	6.100 *	5.946 *	5.787 *
	(1.481)	(1.481)	(3.067)	(1.893)	(2.007)	(1.917)	(1.971)
Percent muslims	0.010 *	0.028 *	0.005	0.005	0.006	0.005	0.006
	(0.005)	(0.005)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
Percent catholics	0.010	0.024 *	0.001	0.002	0.003	0.002	0.002
	(0.005)	(0.005)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percent other religions (Hindu)	-0.000	0.014	-0.010	-0.013	-0.013	-0.012	-0.011
	(0.009)	(0.009)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
1970-1974		11.418 *					
	(1.766)	(1.766)					
1975-1979		11.424 *					
	(1.782)	(1.782)					
1980-1984		13.961 *					
	(1.533)	(1.533)					
1985-1989		14.122 *	15.392 *	16.716 *	16.949 *	17.092 *	16.984 *
	(1.492)	(1.492)	(2.405)	(2.403)	(2.401)	(2.385)	(2.388)
1990-1994		13.393 *	14.463 *	15.707 *	16.783 *	16.535 *	16.908 *
	(1.533)	(1.533)	(2.508)	(2.547)	(2.410)	(2.402)	(2.395)
N	364	364	137	134	134	134	134
R^2	0.577	0.577	0.621	0.614	0.612	0.617	0.616
Adj. R^2	0.559	0.559	0.584	0.576	0.574	0.580	0.579

Standard errors in parentheses
* indicates significance at $p<0.05$

7 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, 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, Strezhnev & Voeten (2013) convincingly highlight that these same measures suffer from agenda effects as resolutions often deal with very topical issues on conflict. Finally, we highlighted in this paper 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 dataset comprising information on allresolutions 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. Neither being particularly friendly with Japan to get bilateral aid in general, nor being friendly with the US in the UNGA to obtain bilateral aid from this country seems to work. In addition, we can show that even when calculating similarity measures only based UNGA decisions deemed important by the US State Department (see Thacker 1999), the results reported by Alesina and Dollar (2000) fail to be robust.

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, have also considerable substantive consquences. Our approach, however, does not deal with the problem highlighted by Bailey, Strezhnev & Voeten (2013), namely possible agenda effects. Their approach to solve this problem can by definition not consider consensus votes and is thus likely to lead to biased estimates as well.

Consequently, future research has to show whether these agenda effects are equally important when considering consensus votes, and how a measurement approach might be developed to address both issues at the same time.

References

Abi-Saab, Georges. 1997. Membership and voting in the United Nations. In *The changing constitution of the United Nations*. London : British Institute of international and comparative law pp. 19–39.

Alesina, Alberto & David Dollar. 2000. "Who Gives Foreign Aid to Whom and Why? " *Journal of Economic Growth* 5(1):33–63.

Alesina, Alberto & Beatrice Weder. 2002. "Do Corrupt Governments Receive Less Foreign Aid?" *American Economic Review* 92(4):1126-1137.

Alker, Hayward R. & Bruce Russett. 1965. *World Politics in the General Assembly*. New Haven: Yale University Press.

Bailey, Michael A., Anton Strezhnev & Erik Voeten. 2013. "Estimating State Preferences from United Nations Data." Paper prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago (April 11-14, 2013).

Blake, Daniel J. & Autumn Lockwood Payton. 2009. "Decision Making in International Organizations: An Interest-Based Approach to Voting Rule Selection." Ohio State University, typescript.

Cassan, Hervé. 1977. "Le consensus dans la pratique des Nations Unies." *Annuaire française de droit international* pp. 456–485.

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. 2000. "Preferences and the Democratic Peace." *International Studies Quarterly* 44(12):191–212.

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.

Hovet, Thomas. 1960. *Bloc Politics in the United Nations*. Cambridge: 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 prepared for presentation at the 5th Conference on The Political Economy of International Organizations, Villanova, January, 2012.

Hug, Simon & Simone Wegmann. 2013. “Ten years in the United Nations: Where does Switzerland stand? ” *Swiss Political Science Review* 19(2): 212–232.

Jessee, Stephen A. 2010. “Issues in Scaling Citizens and Legislator Ideology Together.” The University of Texas at Austin.

Kim, Soo Yeon & Bruce Russett. 1996. “The New Politics of Voting Alignments in the United Nations General Assembly.” *International Organization* 50(4):629–652.

Lewis, Jeffrey B. & Chris Tausanovitch. 2013. “Has Joint Scaling Solved the Achen Objection to Miller and Stokes? ” 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.

Lockwood Payton, Autumn. 2010. “Consensus Procedures for International Organizations.” Max Weber Working Paper Series: EUI MWP 2010/22.

Lockwood Payton, Autumn. 2011. “Building a Consensus (Rule) for International Organizations.” Paper prepared for presentation at the conference The Political Economy of International Organizations, ETH and University Zurich, January 2011.

Mokken, Robert Jan & Frans N. Stokman. 1985. Legislative analysis: methodology for the analysis of groups and coalitions. In *Coalition formation. (Series Advances in psychology, 24).*, ed. H.A.M. Wilke. Amsterdam: North-Holland pp. 173–227.

Rai, Kul B. 1982. “UN Voting Data.” *Journal of Conflict Resolution* 26(1):188–192.

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

Skougarevskiy, Dmitriy. 2012. “Agenda Shift in a Consensus-building United Nations General Assembly.” Graduate Institute of International and Development Studies, Geneva.

Stokman, Frans N. 1977. *Roll calls and sponsorship: a methodological analysis of Third World group formation in the United Nations*. A. W. Sijthoff.

Strezhnev, Anton & Erik Voeten. 2012. “United Nations General Assembly Voting Data.” 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.

Voeten, Erik. 2013. Data and Analyses of Voting in the United Nations General Assembly. In *Routledge Handbook of international organizations*, ed. Robert Reinalda. Routledge pp. 54-66.

dropbox/cvasmi/cvasmi_peio.doc 07/01/14