Extant theoretical literature in International Relations has suggested that the similarity of foreign policy preferences is a meaningful and important concept. The commonly used measures of this concept – $\tau_b$ and, more recently, the $S$-score – are problematic in two ways. First, as algorithms, neither operationalize the concept in a manner consistent with how students of IR theorize foreign policy similarity. Specifically, $\tau_b$ is not a measure of similarity; it is a measure of concordance or rank-order similarity. Although $S$ is a measure of similarity, it does not account for the degree of third-party relationships, only the symmetry of such relationships. A new algorithm, $D$, which incorporates the degree of mutual relationships into the measurement of foreign policy similarity, is introduced. The resulting $D$-score is standardized, as opposed to the $S$-score which has four commonly used versions in addition to several others that may be subjectively defined. Second, the input data for measures of foreign policy similarity are typically data on formal alliances. Generally speaking, alliances are a granular indicator of an underlying, military-strategic relationship between states, more generally referred to as military cooperation. As such, the policy portfolio used to calculate $D$ can include additional indicators of military cooperation. The five derived indicators include alliances, joint military exercises (of which new data has been collected), arms transfers, and multinational combat and peacekeeping operations. The indicators are combined using a graded response model to estimate military cooperation as a latent trait. The latent trait is used as input data to calculate the $D$-score, which is directly compared with the $S$-score and shown to be consistent with intuitive notions of policy similarity while the $S$-score is not. In a replication study, the new measure is shown to not be substitutable with the $S$-score.
Introduction

The United States and Argentina have shared a defensive alliance through the Organization of American States since 1947. The US has had a similar alliance with the United Kingdom through the North Atlantic Treaty Organization since 1949. When the UK went to war with Argentina over the disputed Falkland Islands in 1982, the world wanted to know with whom the US would side.

Answering these sorts of questions is the reason that measures of foreign policy similarity have been developed (Altfeld and Bueno de Mesquita, 1979; Signorino and Ritter, 1999; Hage, 2011). Extant quantitative literature in International Relations (IR) has also shown these measures to be useful for a variety of related theory-testing purposes (Morrow, Siverson and Tabares, 1998; Bapat, 2007; Leeds, 2003; Long and Leeds, 2006; Gartzke, 2007; Moore, 2010). Foreign policy similarity, however, can be interpreted in a variety of ways and there seems to be some confusion as to what concept the accepted measures of policy similarity – current Signorino and Ritter’s $S−score$ and previously Altfeld and Bueno de Mesquita’s $\tau_b$ – are operationalizing.

Bapat (2007) states that “Signorino and Ritter’s weighted measure of $S$ captures the political affinity between target and host states” (p. 276). Morrow, Siverson and Tabares (1998) utilize $\tau_b$ to “operationalize common interests in a dyad of states” (p. 654). In its original formulation, the intent of $\tau_b$ is to measure the extent to which two state’s foreign policy commitments, based on common revealed preferences, are similar (Bueno de Mesquita, 1975; Altfeld and Bueno de Mesquita, 1979). While each of these specifications are related, they are describing a concept represented by neither $S$, $\tau_b$, nor one of Hage’s proposed measures.

$\tau_b$ is a measure of concordance, or rank-order similarity, derived by calculating Kendall’s $\tau_b$ statistic on a matrix containing data on the foreign policies of two states (Bueno de Mesquita, 1975; Altfeld and Bueno de Mesquita, 1979). Despite its widespread use for nearly two decades, concordance is not the similarity of reveal preferences, as Signorino and Ritter (1999) point out. However, $\tau_b$ made two vital contributions to IR that are utilized by its successor measure, the $S−score$. First, seeing as it is constructed from the relationships of actors with every other actor in the system, $\tau_b$ is a network measure.$^1$ Second, $\tau_b$ reflects the intuition that observed foreign policies are revelations of underlying preferences. Each of these contributions is foundational for $S$.

$^1$ $\tau_b$ may be the first network measure used in quantitative IR. Bueno de Mesquita (1975) is, strictly-speaking, a network analysis paper using other terminology.
The $S$ algorithm calculates a geometric distance, commonly the Euclidean distance, between two vectors where each vector corresponds to the foreign policies of two states. As such, $S$ is identical to structural equivalence, a commonly used concept in social network analysis (Lorraine and White, 1971; Wasserman and Faust, 1994). Structural equivalence reflects the symmetry of common relationships in a network and is used to identify actors that share a similar network-position.

In terms of measuring foreign policy similarity, a similar network position does not imply the similarity of revealed preferences. For example, Mexico has a strong relationship with the United States and a relatively weak relationship with most other states in the international system. The same is true for Israel and the Philippines. Although each of these states share a similar network position, not much can be inferred about their political affinity based on this information.

The concern lies in the fact that $S$ fails to capture the degree to which foreign policy commitments are similar. According to the $S$-score, two states with strong common relationships are equivalent to two states with weak or non-existent common relationships. The mutual lack of a relationship is especially troublesome, since two states that have no relationship with any other state have symmetrical foreign policies but are not likely to have any affinity for one another.

I propose a new algorithm, $D$, for calculating foreign policy similarity. $D$ builds upon previous measures by keeping the policy portfolio, network-based approach of $\tau_b$ and the symmetry-based scoring of $S$ but also incorporates the degree of third-party relationships. Using $D$, two states without a third-party relationship are not considered similar, despite the fact that they are symmetrical. Two states with strong third-party relationships upweight the similarity score more than those with weak third-party relationships.

Like the $S$-score, $D$ may be calculated with continuous data and a multidimensional foreign policy portfolio. Traditionally, the foreign policy portfolio has consisted for formal alliances, and students of IR have taken notice of the shortcomings of using a unidimensional portfolio consisting of formal alliances (Bennett and Rupert, 2003; Hage, 2011; Sweeney and Keshk, 2005; Li, 2005; Fearon, 1997; Morrow, 2000). Bueno de Mesquita (1981), Signorino and Ritter (1999), and Gartzke (2007) suggest other potential policies to incorporate, but none have succeeded in supplanting formal alliances.

---

---

$^2$ $S$ also provides the flexibility to include weightings and transformations, both of which are discussed below.
Rather than pushing new policies into the fray, I suggest incorporating those that represent a similar underlying relationship to alliances. Specifically, alliances are a granular indicator of some level of military cooperation among states. Military cooperation is a latent trait of a dyadic relationship and is revealed through a variety of foreign policies including joint military exercises, arms transfers, and multinational combat and peacekeeping operations. The policy portfolio used as input to construct foreign policy similarity, then, is one based on the concept of military cooperation.

Rather than handling a multidimensional policy portfolio through additive indexing, which S does and which entails assuming all policies are equivalent, I suggest using these indicators to estimate military cooperation as a latent trait of a dyadic relationship. Here, this is done using the graded response model (GRM) from item response theory (Samejima, 1997). The $D - score$ is calculated using these estimates.

This article proceeds in three main sections. First, the algorithm used to construct $S$ is discussed in detail and the new algorithm, $D$, is introduced. Second, the primary concerns associated with relying on alliance data and a unidimensional policy portfolio are states, and a new estimate of military cooperation using a GRM is introduced. Third, $D$ is calculated and compared with $S$ and Hage’s measures using data from 1971-2000. Finally, Matthew Moore’s article, Arming the Embargoed, is replicated using the $D - score$ to demonstrate these measures are not substitutable.

**Algorithms**

Foreign policy similarity is not a dyadic measure—it is a network measure whose value depends not only on the dyadic relationship but each member of the dyad’s third-party relationships with every other state in the system. The original motivation for measuring foreign policy similarity grew out of the early expected utility literature where researchers asked with whom a state would be more likely to side with in the event of war (Bueno de Mesquita, 1975, 1981). While similar foreign policy portfolios are meant to be indicative of a strong preference, dissimilar ones are indicative of a weak preference.

Currently, the predominant measure of foreign policy similarity is the $S - score$ (Signorino and Ritter, 1999). Similar to the previous measure of foreign policy similarity, $\tau_b$, the $S - score$ is calculated by performing some algorithm on the foreign policy portfolios of two states. The $S$
algorithm is a distance metric where symmetric portfolios equate to high similarity and asymmetric portfolios to low similarity.

As a measure of the symmetry of two state’s policy portfolios, \( S \) is fundamentally a measure of the social network analysis concept known as structural equivalence; albeit a weighted and transformed version. Deconstructing \( S \) to be directly comparable with a textbook example of structural equivalence is straightforward. Network data are often expressed in a sociomatrix, here denoted \( P \). Suppose we have a two sets of the same actors \( i = 1, \ldots, n \) and \( k = 1, \ldots, m \). The sociomatrix \( P \) is then an \( n \times m \) matrix where each cell contains data about the relation among each pair of actors, \( p_{ik} \). Each row, \( p_i \), refers to the vector corresponding to State \( i \)’s policy portfolio.

According to [Wasserman and Faust 1994, p. 367], structural equivalence is defined as:

\[
d(p_i, p_j) = \sum_{k=1}^{n} [ |p_{ik} - p_{jk}| + |p_{ki} - p_{kj}| ]
\]

where \( n \) is the number of states, \( j \) and \( i \) are policy portfolios, \( i \neq k, j \neq k \), and \( i \neq j \). Since the input data are non-directional, the absolute value of \( p_{ik} - p_{jk} \) equals \( p_{ki} - p_{kj} \) and is therefore equivalent to \( 2 \times |p_{ik} - p_{jk}| \). Equation (1) can be rewritten as:

\[
d(p_i, p_j) = 2 \sum_{k=1}^{n} |p_{ik} - p_{jk}|
\]

As stated above, \( S \) is a weighted and transformed version of structural equivalence as represented by Equation 2. To finish the comparison, let \( w \) equal a vector of weights where \( w_k \) is a weight for each \( k \). In practice, \( w \) is either uniform or corresponds to the Correlates of War’s Composite Index of National Capabilities (CINC) score. If we have uniform weighting, \( w_k \) equals \( \frac{1}{n} \). In the case where \( w \) corresponds to the CINC score, \( w_k \) equals State \( k \)’s CINC score in the given year for which \( S \) is being calculated. By weighting in this way, one could argue that, depending on the relative power of the state, the effect of the third-party on the overall measure of foreign policy similarity is different. 3

Finally, let \( L \) be a vector of scoring rules which maps \( p_i \) to some closed interval along the real number line. \( \Delta_{\text{max}} \) equals the difference between the maximum and minimum possible values after

\[3\] However, note that the choice of weighting scheme is entirely subjective and different weights can have drastically different effects.
the mapping. In practice, \( L \) is not used because alliances, which are integer values typically not reaching a value higher than 4, are quite easy to work with.

As is the case in practice, we will look at the equation for calculating \( S \) for a single policy dimension where the sum of the weighting vector equals 1 and we assume that no mapping for the scoring rules is necessary. Signorino and Ritter’s original series of equations (p. 127) can then be rewritten as:

\[
S(p_i, p_j, W, L) = 1 - 2 \left( \frac{1}{\Delta_{\text{max}}} \right) \sum_{k=1}^{n} (w_{k,k} |p_{ik} - p_{jk}|)
\] (3)

Unfortunately, structural equivalence is not the concept we want when measuring foreign policy similarity because symmetry is not affinity. To demonstrate this point graphically, the \( S \) algorithm is visualized in Figure 1.

In Figure 1 we have a set of six states, \{A, B, C, D, E, F\}, and are calculating the \( S \)-score for States \( i \) and \( j \) using their mutual relationships with these six third-party states. \(^{4}\) State \( A \), for example, scores a 1 for its relationship with both State \( i \) and State \( j \). State \( C \) scores a 4 with State \( i \) and a 1 with State \( j \).

Because \( S \) measures the symmetry of common relationships, all points on the diagonal – States \( D, A, \) and \( B \) – represent the same contribution to the dyad’s \( S \)-score. Because all points on the diagonal are perfectly symmetrical, they also represent the maximum contribution for any individual third-party relationship. For all points off the diagonal, as with States \( E, F, \) and \( C \), each point’s contribution to the \( S \)-score is the horizontal (or vertical) distance from the point to the diagonal, illustrated by the red dashed line. Essentially, any point on the diagonal contributes to perfect similarity while points off the diagonal reduce the similarity score.

The disconnect between the operationalization of our measure of foreign policy similarity and our conceptualization of foreign policy similarity can be easily highlighted with a couple examples from Figure 1. To better highlight the differences, assume the points represent an alliance category from the Alliance Treaty Obligations and Provisions (ATOP) dataset \(^{5}\) (Leeds et al., 2002). According to this schema, 1 equals a non-aggression pact, 2 equals neutrality agreement, 3 equals a defensive alliance, and 4 equals an offensive alliance.

---

\(^{4}\)This figure is an accurate depiction of \( S \) in the case of uniform weights, no weights, or when the weights are state and not dyad specific.
First, compare the different effects of States \( F \) and \( D \). Because the relationship is not symmetrical, State \( F \) will act to reduce the foreign policy similarity of States \( i \) and \( j \) while State \( D \), because the relationship is symmetrical, contributes to perfect similarity. By not incorporating magnitude into the measure we disregard the fact that State \( i \) maintains a defense pact with State \( F \) and State \( j \) maintains an offensive alliance with State \( F \) – both indicative of strong preferences.

Second, compare the effects of States \( D \), \( A \), and \( B \) on the measure of foreign policy similarity. Each relationship with States \( i \) and \( j \) is symmetrical and therefore falls on the diagonal. Because each relationship is symmetrical, all contribute identically to perfect similarity. This is the case despite the fact that States \( i \) and \( j \) have no relationship with \( D \) and a defense pact with \( B \). Once again, a defense pact is indicative of a strong preference but, by not incorporating magnitude, the \( S - score \) does not reflect this.

A more appropriate measure of foreign policy similarity should not only incorporate symmetrical relationships but also the magnitude of the third-party tie. For instance, as states move further away from the origin along the diagonal, they should contribute more; State \( B > State \ A > State \ D \). Holding symmetry constant, a greater degree of magnitude reveals a stronger preference.

When symmetry is not present, as with State \( F \), incorporating magnitude is more subjective. How much should the similarity score be reduced for the off-diagonal component? Mathematically,
there are many ways to combine magnitude and symmetry to reflect a single value for the mutual relationship. While the best approach is therefore debatable, a conservative estimate is to use the weak link assumption, which is to say that the contribution to the overall score is the lesser of the two relationships with the third party.

The $D$ algorithm incorporates these principles and is shown in visually in Figure 2. For all points that fall above the diagonal, that point’s contribution equals the horizontal distance from the point to the Y-axis (e.g. State $F$). For all points below the diagonal, the contribution equals the vertical distance to the X-axis (e.g. State $C$). For all points on the diagonal, the contribution equals either the horizontal distance to the Y-axis or the vertical distance to the X-axis (here shown as the horizontal distance for States $A$ and $B$).

As can be seen in the dashed lines in Figures 1 and 2, which represent the contribution to the scores from each third-party, $S$ and $D$ are structurally different. For example, States $F$ and $B$ contribute quite differently depending on which measure is used. $S$ is not a measure of magnitude, and so State $B$ has the same contribution as States $A$ and $D$. Using $D$, both States $F$ and $B$ contribute the same amount to the measure of policy similarity; a contribution which is larger than that of other states.

Formally, the basic calculation of $D$ is simple. Using the same notation as above:

$$D(p_i, p_j) = \sum_{k=1}^{n} \min(p_{ik}, p_{jk})$$ (4)

[Hage (2011)] proposes to apply Scott’s $\pi$ and Cohen’s $\kappa$, two statistics that incorporate chance correction to the measurement of foreign policy similarity.\footnote{Cohen’s $\kappa$ also accounts for the fact that some states will exhibit greater propensities to form alliances than others.} Chance correction allows for the possibility that both states will have a mutual relationship with a third-party by chance. This is called a chance agreement. More specifically, if the foreign policy portfolios were randomized, a chance agreement occurs when both states randomly have an identical policy towards a third-party (Gwet, 2002). Given the sparsity of foreign policy portfolios, chance correction mostly applies to the non-existent third-party relationships.

Although Hage’s measures are an improvement over the $S – score$ in general applications, they still does not account for the degree of mutual relationships. Common third-party relationships
are still equivalent, regardless of how strong the relationship actually is. Dissimilar but strong third party relationships are still dissimilar and will penalize the foreign policy similarity score. The $D – score$, in contrast, accounts for these shortcomings. Furthermore, Scott’s $\pi$ and Cohen’s $\kappa$ require interval level data, an assumption causes problems for the multidimensional portfolio approach.

Cooperation and Revealed Preferences

In addition to its aforementioned shortcomings, the $S – score$ induces a lack of standardization into the measurement of foreign policy similarity. As specified by Signorino and Ritter (1999), weighting schemes, scoring rules, the distance metric, and the composition of the portfolio are all subjective decisions. Furthermore, the $S – score$ has also been modified to incorporate regional and global distinctions in policy portfolios (Bennett and Stam, 2000). This lack of standardization is problematic for comparison.

The $D – score$ does not have weighting schemes or scoring rules. Because the lack of a relationship (zeros in the policy portfolio) does not artificially inflate the similarity score, it is not necessary to calculate anything regionally. Furthermore, the algorithm specifies the lesser of two values be used in each pairwise comparison, so there is no decision as to which distance metric is to be used. However, the construction of the $D – score$ does entail one moving part: the foreign policy portfolio.

The policy portfolio used to construct measures of foreign policy similarity has consisted solely of formal alliances. For measuring foreign policy similarity, alliances have their share of shortcomings, including the fact that having no alliance can mean multiple things, states do not always choose to ally, and states value alliances with other states differently (Wagner, 1984; Bennett and Rupert, 2003; Sweeney and Keshk, 2005; Li, 2005; Fearon, 1997; Morrow, 2000). Specifically, from 1971 - 2001 there are 428,015 dyad-years in the data and 399,071 have no alliance. Of these, the “no alliance” category includes “those that have no alliance with each other because they are hostile to each other, those that have no alliance because they are irrelevant to each others security, and those that have no alliance because of implicit alignment that renders a formal treaty unnecessary”

These numbers are calculated using the COW alliance data and state-system membership.
Each of these relationships are identical if our sole indicator is alliances.

Furthermore, states value their alliances differently. Mexico and Uruguay are defensive allies—so are the UK and France. These dyadic relationships are not identical, but they appear to be so through the lens of a formal alliance. The difference can be seen when expanding the foreign policy portfolio to include other valid policies. Specifically, since alliances have survived since 1979 as a useful indicator, I ask what other policies are indicative of the same underlying relationship as alliances.

Alliances are viewed as one indicator of a dyad’s latent level of military cooperation. Other indicators of dyadic military cooperation include joint military exercises, arms transfers, and multinational peacekeeping and combat operations. Each of these four indicators plus alliances are used to estimate military cooperation via the graded response model (Samejima, 1997).

**Policy Portfolio Data**

Military cooperation is defined to take place when a policy is adopted by two or more states for the purposes of either (1) reducing one or more excluded state’s military power, (2) enhancing the military power of at least one state, or (3) coordinating military activity for the benefit of some state or set of states. The policy must deliberately involve a military in some capacity, and may be either behavioral or institutional.

At the dyadic level, military cooperation is an underlying, latent and continuous concept that manifests in various observable ways. Each of these manifestations is an indicator of military cooperation. The five indicators used here in the policy portfolio are briefly discussed below.

**Alliances**

Despite their granularity, alliances are a significant indicator of military cooperation. Leeds et al. (2002) define alliances as “written agreements, signed by official representatives of at least two independent states, that include promises to aid a partner in the event of military conflict, to remain neutral in the event of conflict, to refrain from military conflict with one another, or to

---

7This section is very similar if not identical to the description in Chapter Four.
consult/cooperate in the event of international crises that create a potential for military conflict” (p. 238). In essence, alliances are a pledge to act in a certain way in the event of war.

Alliances can signal various depths of cooperation, depending on the actions that a state pledges to take in the event of war. The Correlates of War data, which is used here, defines three types of alliances, which are ordered in terms of the depth of cooperation entailed (Gibler and Sarkees, 2004; Gibler, 2009). A defensive alliance, which signals the deepest cooperation, requires signatories to come to the defense of one another should one be attacked. Non-aggression and neutrality agreements are each coded at the middle level of cooperation, but they are also slightly different from one another. A non-aggression pact ensures that the signatories will not declare war on one another while a neutrality pact requires states to at least remain neutral in the event of war. Finally, an entente is a pledge to consult one another should a conflict emerge. Here, an entente is coded as the weakest form of alliance.

**Arms Transfers**

Similar to an alliance, arms transfers are a security-enhancing indicator of military cooperation. They, too, reveal a depth of cooperation, both in the simple decision to deal arms to a state as well as in the amount of arms dealt to a state. As such, arms transfers are more than just sale for profit. They are a strategic policy option meant to shore up a relationship. States do not deal arms to those who they are not friendly with for fear that those weapons could be turned against them.

Data on arms transfers is made available by the Stockholm International Peace Research Institute (SIPRI). For every year, the value of the arms transfer is scored using SIPRI's trend indicator value (TIV). The TIV is SIPRI's valuation of the arms transfer based in constant 1990 U.S. dollars. According to Dr. Holtom, “SIPRI TIV are best used as the raw data for calculating trends in international arms transfers over periods of time, global percentages for suppliers and recipients, and percentages for the volume of transfers to or from particular states” (personal communication, December 9, 2011). As a result, the TIV is a continuous variable and it is assumed that the greater the value of the arms transfers, the greater the depth of cooperation.
Joint Military Exercises

A third type of military cooperation comes in the form of joint military exercises (JMEs). JMEs, or war games, take place when the militaries from more than one state interact in such a way as to enhance their ability to carry out military operations. Although they have a wide range of practical purposes, JMEs also have political implications ranging from boosting a state’s status to disguising military buildups. Different types of exercises serve different purposes; while some are meant to signal a commitment or a willingness to get involved in a conflict or to deter states from further escalations towards conflict, others may be held primarily as a coordination exercise among members of the same international organization.

JMEs satisfy the second criteria for military cooperation in that they enhance the security of at least one, and in this case both, participating states by improving the military’s ability to carry out military operations. Although different types of exercises reveal different levels of cooperation, due to time constraints the magnitude of the exercise has not been coded. Therefore, depth of cooperation for this indicator is measured by the number of JMEs a dyad participates in for a given year. The data on JMEs, as well as comparisons to the data on alliances and arms transfers, is discussed in more detail in Chapter Three.

Multinational Military Operations

A fourth type of military cooperation comes in the form of multinational military operations. Within this category are both combat and peacekeeping operations. Similar to a military exercise, these are behavioral forms of cooperation. In the case of combat operations, multiple states are cooperating to reduce the security of another state by carrying out a military operation against that state. In the case of peacekeeping operations, multiples states are promoting the security of another state by executing a military operation designed to maintain or enforce peace.

The multinational combat operations indicator has been constructed using PRIO’s armed conflict data. All cases of more than one state fighting on the same side of an issue are coded as those states cooperating in a combat operation. Zeev Maoz’s dyadic MID data (Ghosn, Palmer and Bremer, 2004; Maoz, 2005) present another option for operationalizing the multinational combat operations indicator. The MID data are much less restrictive than the PRIO data because they
do not require battle-deaths to have occurred. However, the MID data are also limited to international conflict where the PRIO data contain both international and intra-state conflict. Ideally, it is possible for both indicators to be included, but at present the MID data do not have sufficient temporal coverage.

The peacekeeping data have been coded from the set of peacekeeping operations that have been labeled by Mark Mullenbach as part of the Dynamic Analysis of Dispute Management project. These data include both state-led and organization-led operations, but are limited to peacekeeping operations in intra-state conflicts. Although there are other mechanisms for capturing the level of participation by a state in a given multinational military operation, here the depth of cooperation is measured by the number of joint operations in a given year.

Other Indicators of Military Cooperation

Other indicators of military cooperation could also be included when measuring this concept. Possibilities of future event-based indicators include military aid, some specific types of sanctions, and international military education. Future structural indicators could include membership in regional security institutions, such as the Shanghai Cooperation Organization, and general defense cooperation and information security agreements (Kinne, 2012). However, due to a lack of data, these indicators have not been included.

The Military Cooperation Estimate

While the $S$ algorithm allows for multiple policies in each state’s portfolio, the assumptions required are unrealistic. $S$ handles multiple policies by constructing an additive index from each policy, which means that all policies are assumed to be of equal weight and a one-unit increase in one policy is identical to a one-unit increase in any other. Furthermore, it also assumes that increases of $x$-units on any individual policy are identical regardless of where the increase takes place.

Each of these assumptions are unrealistic, since policies are not as directly comparable as $S$ would hope. Fortunately, none of those assumptions are necessary when using the GRM to estimate military cooperation as a latent trait. Rather than assuming each indicator is weighted the same,
the GRM estimates the effect of the latent trait on the indicators with respect to one another. Low
levels of military cooperation, for example, might result in an observed multinational peacekeeping
operation but not in a defensive alliance, which may be reserved for those observations with a
higher level of military cooperation.

For the GRM, the effects of increases within an indicator are also estimated as a function of
the latent trait so that an x-unit change in the indicator does not always correspond with a y-unit
change in the estimated military cooperation. Instead, the affect on the measure is the result of

where

the x-unit change in the indicator occurred.

The item responses are assumed to be ordinal, which is the case with each of the five indicators
except for arms transfers. The arms transfer data, SIPRI’s TIV, is continuous and the graded
response model requires ordinal data. Therefore, it is scaled to represent five levels of arms transfer.

The first is a level of 0 where no arms transfer has occurred. The vast majority of dyad-years from
1971 through 2000 – 401,004 out of 409,872 – fall into this category. The remaining 8,868 dyad-
years are split roughly evenly into the remaining four levels, with 2,172 in the first, 2,124 in the
second, 2,216 in the third, and 2,354 in the fourth. Similarly, the JME and peacekeeping indicators
range from zero to eleven and zero to twelve, respectively, and are based on the number of JMEs
or peacekeeping operations that both states participated in for a given year.

Treier and Jackman (2008) use a similar method for calculating a regime-type measure using
the Polity indicators, and Benson and Clinton (2012) use an application of item response theory
to measure the strength of an alliance across two dimensions: the power of the individual states that
are party to the agreement and the terms of the agreement itself. The notation here is similar to
that applied by Treier and Jackman (2008). Specifically, each dyad-year is indexed by \( i = 1, \ldots, n \).
The indicators of military cooperation are represented by \( j = 1, \ldots, m \) and the set of ordinal responses
for each \( j \) is given by \( k = 1, \ldots, K_j \). The following model is estimated using the \texttt{grm} function in the
\texttt{ltm} package in R (Rizopoulos 2006).

\[ \text{The choice of five is arbitrary.} \]
\[
P(y_{ij} = k|x_i) = g(\eta_{jk}) - g(\eta_{j,k+1}),
\]
\[
\eta_{jk} = \alpha_j(x_i - \beta_{jk}), \quad k = 1, ..., K_j,
\]
\[
g(z) = \frac{1}{1 + e^{-z}}
\]

The model estimates a dyad’s latent level of military cooperation for a given year. In comparison to alliances, this estimate provides a more complete measure of the underlying dyadic relationship necessary to measure foreign policy similarity. To provide some face validity for this measure, therefore, it is compared with alliance data for the United States. To do so, three alliance categories are subset from the data: no alliance, an entente, neutrality, or non-aggression pact, and defense pacts. Each of these three types are further split into three decades: 1971-80, 1981-90, and 1991-2000.

For each decade and for each alliance type, the estimate of military cooperation with the United States for a state in a given year is placed on a scale from one to twenty. This scale is divided into equal parts based on the range of possible estimates of military cooperation. Although twenty is an arbitrary number, the distributions in Figure 3 serve the primary point.

Figure 3 demonstrates that considerable variation within alliance categories exists for many of the dyad-years in the data. Additionally, there are two important trends to point out. First, the variance within any alliance category is increasing over time. This is seen by the increasing skew of the distribution that is observed in each of the alliance types as one moves across decades. This means that states are increasingly engaging in more types of military cooperation with the United States than they have in the past. Second, there is variation across alliance categories. There are many states that the United States cooperates with but does not have an alliance with. Israel is the most commonly used example of such a relationship. In other cases, a defense pact is present, but no other indicator of military cooperation is present. This is represented by the lowest bar in each of the three Defense Pact plots in Figure 3.

The granularity of alliance data make it inappropriate to use as the sole indicator of revealed foreign policy preferences. Other forms of military cooperation are indicative of those same preferences and may be used as part of a state’s policy portfolio. Their incorporation into a measure
Figure 3: United States
of foreign policy similarity, however, ought to be as thoughtful as their inclusion into the policy portfolio. Simple additive indexing is an assumption-laden approach, and better methodologies for combining indicators exist. The GRM is one such method, but certainly others exist and may even provide more intuitive estimates.

Thus far I have suggested a new algorithm for calculating foreign policy similarity and a new approach towards the foreign policy portfolio. The following section compares the resulting $D$–score to the $S$–score and to Hage’s measures.

**Empirical Comparisons**

The $S$–score lacks standardization because of the subjective decisions entailed in its construction. As a result, four mainstream $S$–scores exist: unweighted global, weighted global, unweighted regional, and weighted regional. The distributions of these scores, shown in Figure 4, demonstrate precisely why standardization is desirable.

Theoretically, the $S$–score ranges from -1 to 1. Empirically, however, the range depends largely on the choice of which $S$–score is used. The unweighted global score ranges from 0.056 to 1, with the majority of observations scored greater than 0.5. The weighted global $S$–score ranges from -0.525 to 1, but again the vast majority of mass is at or greater than a score of 0.5. Although the range of the distribution of the weighted regional $S$–score is from -0.984 to 1, by far most of these observations are scored at or near a perfect similarity score of 1. The unweighted regional score, which has a range of -0.984 to 1 and contains significant mass at both ends of that spectrum, is perhaps more reasonable given our expectations of foreign policy similarity, but there still appears to be many dyad-years that are extremely similar. Given their varying distributions, no replication is necessary to demonstrate these measures are not substitutable, leaving researchers to question which is the most appropriate.

Figure 4 appears so skewed because of the many relationships in the international system that are symmetrical due to the lack of a relationship. Hage (2011) proposes to use Cohen’s $\kappa$ or Scott’s $\pi$ to account for this sparsity. As shown in Figure 5, the distribution of Hage’s two measures are quite similar to one another. However, they are quite different from the $S$–scores, and instead

---

10These $S$–scores are available through EUGene until 2000. The estimates of military cooperation begin in 1971, and so the overlapping period is 1971 through 2000.
of having the most mass at -1 or 1 they have most mass at 0. The $\kappa$ and $\pi$ measures range from -0.293 and -0.351 to 1, respectively.

It is interesting that there is so little mass from scores of about 0.1 up until 1, and then significant numbers of observations at 1. Specifically, there are 42,073 observations of 1 out of a total 409,872 observations in the data. Upon closer look, these states are mostly those of NATO and the OAS, although there are other dyads in there as well. It is a feature of these chance-corrected methods that several dyads will be coded at perfect similarity. By construction, the $D - score$ is quite different.

For the $D - score$, most observations are scored at or near -1 and there is a long tail to the distribution. By construction, the $D - score$ has a minimum of -1 and a maximum of 1, and each will be represented by construction. However, only the highest scoring observation is scaled to equal 1, and for most years this will constitute only one observation. This is in contrast to previous measures where perfect similarity is frequently observed.

Thus far I have argued that the method proposed here improves both on previous algorithms and on previous foreign policy portfolio constructions. To demonstrate the differences in $D$ and $S$ for foreign policy similarity, both scores are plotted side by side with one another using the estimated military cooperation as input and alliances as input.

The first set of plots are the $D$ and $S$ scores using military cooperation and alliances for the United States with Great Britain, India, Israel, Argentina, Russia, and Mexico from 1971 through 2000. Holding the algorithms constant, the construction of the foreign policy portfolio by estimating military cooperation as a latent trait adds considerable variation to the measure. This is highlighted in the difference when moving from the bottom row of figures to the top. Similarly, there is added variation depending on the algorithm. This is highlighted when moving from the right column (the $S$ algorithm) to the left column (the $D$ algorithm).

The first case to point out is the difference in Argentina’s and the UK’s foreign policy similarity with the US. The UK and Argentina went to war in 1982 over the disputed Falkland Islands and the world was curious with whom the US would side. Eventually, the United States supported the UK with military equipment but did not directly intervene. With the exception of Figure 7, all constructs of foreign policy similarity plotted here got this wrong and placed Argentina above the UK. Figure 7 uses the $D$ algorithm along with the estimated military cooperation through the
Figure 4: Four S-Scores: 1971 – 2000
Figure 5: Hage’s Kappa and Pi: 1971 – 2000

Figure 6: The D-Score: 1971 – 2000
multidimensional policy approach.

The divergence of Mexico and Argentina beginning in the late 1980s in Figure 7 is also consistent with history and not plotted in any other figure. Despite both being members of the Organization of American States, which means they have a similar set of alliances, Mexico has not integrated its foreign policies the same way that Argentina has. Argentina is active in peacekeeping operations, joint military exercises, and multinational combat operations. Generally speaking, Argentina is a regular for military cooperation with the states of Western Europe. Mexico, on the other hand, is largely preoccupied with domestic concerns and has not taken its military to the international system in the same way that Argentina has.

Anecdotally, the divergence of Mexico and Argentina can be explained by increased military cooperation. However, Figure 8 which uses the $S$ algorithm on the estimated military cooperation, registers hardly any difference in Mexico and Argentina until about 1989. It certainly does not reflect the spike in military cooperation seen in Figure 7 in the late 1970s and early 1980s. Furthermore, when the divergence does begin, it is quite marginal, which is consistent with the $S$ algorithm since the $S$ score does not empirically range from -1 to 1.

![Figure 7: D(military cooperation)](image)

![Figure 8: S(military cooperation)](image)

Other states also reflect history more accurately in Figure 7 than in others. For example, there is an increase in foreign policy similarity beginning around 1990 with respect to Russia and India,

\[11\] The range of the $S$ score calculated using military cooperation is from 0.084 to 0.996.
which should not come as a surprise considering the way the international system has changed since the end of the Cold War. Military cooperation in particular has been increasing, and it should therefore be expected that a measure of security policy similarity reflect such increases.

Figure 9 demonstrates some of the fundamental issues with using the $S$-score. Specifically, the $S$-score is universally inflated and does not change much over time. The $D$-score, as calculated with alliance data in Figure 9, also does not change much over time. However, the similarity scores with Russia, Israel, and India are all extremely low, which is what one would expect when calculating the foreign policy similarity for states that have very few formal alliances in common. Contrast this to the policy similarity scores seen in Figure 10 and another primary difference in the two measures is seen. The $S$-score inflates these values while the $D$-score values them at or near zero.

Russia’s foreign policy similarity, calculated from 1971-2000 with military cooperation, is plotted for Great Britain, India, Israel, Argentina, the United States, and Poland. The $S$-score again reflects a largely constant measure while the $D$-score reveals a different trend; one that is a more appropriate fit for how recent Russian history has unfolded. In Figure 11, the sharp increase in policy similarity after the Cold War is what would be expected for Russia. The United States, United Kingdom, and Poland are the three most similar states in Figure 11 with India and Argentina having a roughly identical similarity to Russia and Israel is again at or near the lowest.
Israel and Poland reflect perhaps the most interesting types of relationships here. For the $D - score$, whether it be with the United States or Russia, the security policy similarity with Israel is at-or-near the lowest. This reflects the empirical fact that Israel does not cooperate militarily with many states; its security policies are considerably more local. Contrast this with the United States, whose security policy is global.

Poland represents another particularly interesting example. During the Cold War, Russia and Poland were close allies with very similar foreign policies. Since the Cold War ended, Poland’s foreign policy has been largely in line with Western interests. Poland joined NATO, for example, and was one of just three other states to initially contribute troops to the United States’ 2003 invasion of Iraq. The $D - score$ shows a level of policy similarity just below that of the US and UK while the $S - score$ puts Russia and Poland near one.

Figures 13 and 14 show Hage’s constructs of foreign policy similarity among Russia and the same set of states. Both measures are very similar, and both capture Russia’s relationship with Poland better than either the $D - score$ or the $S - score$. However, both also fail to capture any meaningful or sustained change in similarity among Russia and either US or UK. Historically, we know that when the Cold War ended Russia moved considerably closer to Western Europe than it had been during the Cold War, and despite a brief increase in similarity around 1990 this is not reflected.
Replication Study

Matthew Moore’s 2010 article in the *Journal of Conflict Resolution* is an example of an article that uses $S$ as its main explanatory variable. Moore tests a supply-side theory of arms embargo violations and finds that, provided a state is going to violate an embargo, it will do so to a greater degree with a state that has similar foreign policies to that of the violator. Here, I replicate those results and demonstrate that $S$ and $D$ are not substitutable with one another.

The results from Moore (2010) have been replicated in Table 1. Although Moore does not find statistically significant results for $S$ in this experiment, he does find $S$ to have a positive coefficient, which is what he expects. In the replicated results using $D$, the coefficient is actually negative, although it is quite close to 0.

Moore’s main findings are replicated in Table 2. In this analysis, there are only 89 observations because he subsets the data based on which states have actually violated arms embargoes. After subsetting, my data contain 89 observations as well and his results are perfectly replicated. The key finding in Moore (2010) is that $S$ is statistically significant and positive, with the interpretation being that, provided a state is going to violate an arms embargo, it is going to do so to a greater degree.

---

12 Although I have not been able to perfectly replicate the results, they are quite close. Notice the observations go from 872 to 878; this is because of a merging issue that I have to deal with. Once I find out which six observations were either duplicated or included by accident, I believe I will be able to replicate Moore’s results exactly. Neither the coefficients nor the standard errors are off by much; typically just a couple hundredths. All the coefficients are statistically significant when they should be and with a common level of confidence.
Table 1: Replicating Moore’s Table 2

<table>
<thead>
<tr>
<th></th>
<th>Original Results</th>
<th>Replicated</th>
<th>With $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$ score</td>
<td>0.624 (0.680)</td>
<td>0.636 (0.667)</td>
<td>–</td>
</tr>
<tr>
<td>$D$ score</td>
<td>–</td>
<td>–</td>
<td>-0.001 (0.009)</td>
</tr>
<tr>
<td>Arms import dependence</td>
<td>1.583** (0.641)</td>
<td>1.598** (0.686)</td>
<td>1.86** (0.707)</td>
</tr>
<tr>
<td>Military expenditure</td>
<td>0.240* (0.123)</td>
<td>0.241* (0.131)</td>
<td>0.228* (0.123)</td>
</tr>
<tr>
<td>Total annual arms imports</td>
<td>0.054 (0.0882)</td>
<td>0.054 (0.090)</td>
<td>0.058 (0.090)</td>
</tr>
<tr>
<td>Arms export dependence</td>
<td>-1.902 (1.816)</td>
<td>-1.859 (1.695)</td>
<td>-1.45 (1.632)</td>
</tr>
<tr>
<td># of arms transfer relationships</td>
<td>0.078 (0.107)</td>
<td>0.084 (0.106)</td>
<td>0.080 (0.108)</td>
</tr>
<tr>
<td>Voluntary embargo</td>
<td>0.706** (0.278)</td>
<td>0.705** (0.311)</td>
<td>0.736** (0.293)</td>
</tr>
<tr>
<td>Comprehensive sanctions</td>
<td>-1.185*** (0.278)</td>
<td>-1.185*** (0.295)</td>
<td>-1.189*** (0.291)</td>
</tr>
<tr>
<td>Cold War</td>
<td>0.241 (0.340)</td>
<td>0.249 (0.333)</td>
<td>0.266 (0.334)</td>
</tr>
<tr>
<td>South Africa</td>
<td>0.617 (0.458)</td>
<td>0.571 (0.444)</td>
<td>0.561 (0.449)</td>
</tr>
<tr>
<td>Constant</td>
<td>-5.970*** (1.833)</td>
<td>-6.022*** (1.960)</td>
<td>-5.375*** (1.735)</td>
</tr>
<tr>
<td>Observations</td>
<td>872</td>
<td>878</td>
<td>878</td>
</tr>
</tbody>
</table>

degree with an embargoed state whose foreign policies are similar to that of the violating state. When substituted, $D$, however, is statistically significant and negative. Of course, this leads to the exact opposite interpretation and a different view of embargo violators.

By contrast, Hage’s $\kappa$ measure supports the original findings when using the $S$ – score. That is, provided a state is going to violate an arms embargo, it will do so with in greater quantities with a state whose foreign policy is similar to its own. The magnitude of the coefficient is much larger for $\kappa$ as well, which may lead to different substantive effects.

**Conclusion**

Foreign policy similarity has been a variable used by students of IR for over thirty years. Altfeld and Bueno de Mesquita (1979) used a network-based approach and made the argument for using alliances as the primary indicator. The resulting measure, $\tau_b$, was used to proxy for foreign policy similarity for about 20 years. In 1999, Signorino and Ritter introduced a new algorithm for measuring foreign policy similarity. The $S$ – score retains the network-based approach of $\tau_b$ and uses alliances as the primary indicator. However, it is different from $\tau_b$ in that it is a measure of similarity while $\tau_b$ is not. Hage (2011) has introduced the idea of accounting for randomly identical responses, which presents a much more accurate distribution of foreign policy similarity than that of $S$. 
Table 2: Replicating Moore’s Table 3

<table>
<thead>
<tr>
<th></th>
<th>Original Results</th>
<th>Replicated</th>
<th>With D</th>
<th>With Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>S score</td>
<td>2.670** (1.189)</td>
<td>2.670** (1.189)</td>
<td>-0.019* (0.011)</td>
<td>-</td>
</tr>
<tr>
<td>D score</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Hage’s κ</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>16.922* (8.979)</td>
</tr>
<tr>
<td>Arms import dep</td>
<td>-0.650 (0.769)</td>
<td>-0.650 (0.769)</td>
<td>-0.769 (0.781)</td>
<td>-0.788 (0.749)</td>
</tr>
<tr>
<td>Military expenditure</td>
<td>-0.488 (0.362)</td>
<td>-0.488 (0.362)</td>
<td>-0.728* (0.366)</td>
<td>0.684 (0.540)</td>
</tr>
<tr>
<td>Tot ann arms imports</td>
<td>0.080 (0.144)</td>
<td>0.080 (0.143)</td>
<td>0.123 (0.146)</td>
<td>-0.013 (0.160)</td>
</tr>
<tr>
<td>Arms export dep</td>
<td>-0.976 (2.042)</td>
<td>-0.976 (2.042)</td>
<td>1.444 (1.692)</td>
<td>0.416 (1.736)</td>
</tr>
<tr>
<td># arms trans rels</td>
<td>0.130 (0.178)</td>
<td>0.130 (0.178)</td>
<td>0.094 (0.183)</td>
<td>0.204 (0.175)</td>
</tr>
<tr>
<td>Voluntary embargo</td>
<td>0.460 (0.888)</td>
<td>0.460 (0.888)</td>
<td>-0.244 (0.862)</td>
<td>3.538* (1.452)</td>
</tr>
<tr>
<td>Comprehensive sancs</td>
<td>-1.402 (1.273)</td>
<td>-1.402 (1.273)</td>
<td>-0.971 (1.291)</td>
<td>-0.311 (1.285)</td>
</tr>
<tr>
<td>Cold War</td>
<td>-0.663 (0.520)</td>
<td>-0.663 (0.520)</td>
<td>-0.772 (0.528)</td>
<td>-0.100 (0.551)</td>
</tr>
<tr>
<td>South Africa</td>
<td>0.806 (0.844)</td>
<td>0.806 (0.844)</td>
<td>0.565 (0.844)</td>
<td>0.053 (0.834)</td>
</tr>
<tr>
<td>Lagged Volume</td>
<td>0.218*** (0.0975)</td>
<td>0.218** (0.097)</td>
<td>0.217** (0.099)</td>
<td>0.287** (0.100)</td>
</tr>
<tr>
<td>UKR-LBY 2001</td>
<td>3.745*** (1.327)</td>
<td>3.745*** (1.327)</td>
<td>3.167** (1.341)</td>
<td>NA</td>
</tr>
<tr>
<td>ISR-ZAF 1986</td>
<td>2.899** (1.221)</td>
<td>2.899** (1.221)</td>
<td>2.922** (1.238)</td>
<td>3.242** (1.201)</td>
</tr>
<tr>
<td>ISR-ZAF 1987</td>
<td>-3.630*** (1.231)</td>
<td>-3.630** (1.231)</td>
<td>-3.711*** (1.247)</td>
<td>-3.611** (1.203)</td>
</tr>
<tr>
<td>ERI-SOM 1999</td>
<td>-4.469** (1.929)</td>
<td>-4.469** (1.929)</td>
<td>-5.68*** (1.967)</td>
<td>0.264 (2.552)</td>
</tr>
<tr>
<td>Constant</td>
<td>6.144 (5.539)</td>
<td>6.144 (5.539)</td>
<td>12.020** (5.375)</td>
<td>-8.743 (7.932)</td>
</tr>
<tr>
<td>Observations</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
</tr>
</tbody>
</table>

In this article I have shown that the method for constructing $S$ is actually a method for measuring structural equivalence, a concept from social network analysis that is used to measure the similarity of positions of two nodes in a network. While potentially of importance to the study of IR, structural equivalence is not how we should be measuring foreign policy similarity because it does not incorporate the magnitude of common relationships, only their symmetry.

A new algorithm for constructing foreign policy similarity, $D$, is introduced. $D$ incorporates the network-based approach of $τ_b$, as well as including aspects of symmetrical relationships into the measure. As an improvement upon $S$, $D$ incorporates the degree of common relationships.

Additionally, instead of relying just on alliances as the sole indicator of foreign policy, the $D$ – score includes multiple indicators of the more general, underlying relationship of military cooperation. By estimating military cooperation as a latent relationship using item response theory, the indicators of military cooperation can be combined into a single measure to then use as input data for the $D$ algorithm.

Each of the measures explored here, the $S$ – score, Hage’s $κ$ and $π$ measures, and the $D$ – score, have clear differences from one another. This has been evidenced by differences in their theoretical...
foundations, descriptive comparisons, and by their different effects in a replication study. Moving forward, to better understand which measure to use scholars should precisely define their concepts and select the measure that most accurately reflects such concepts.
References


Gwet, Kilem. 2002. “Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters.”.


