A Latent Variable Approach to Measuring and Explaining Peace Agreement Strength*

Rob Williams, Daniel J. Gustafson, Stephen E. Gent, and Mark J.C. Crescenzi

Much of the peace agreement durability literature assumes that stronger peace agreements are more likely to survive the trials of the post-conflict environment. This work does an excellent job identifying which provisions indicate that agreements are more likely to endure. However, there is no widely accepted way to directly measure the strength of agreements, and existing measures suffer from a lack of nuance or reliance on subjective weighting. We use a Bayesian item response theory model to develop a principled measure of the latent strength of peace agreements in civil conflicts from 1975-2005. We illustrate the measure's utility by exploring how various international factors such as sanctions and mediation contribute to the strength or weakness of agreements.

Introduction

The study of civil conflict resolution is rife with weak peace agreements that were unable to bring closure to their respective conflicts. The Arusha Accords, signed in 1993 to end a three-year Rwandan Civil War, infamously failed to prevent the recurrence of conflict in

*Rob Williams (jrw@live.unc.edu) Ph.D. Candidate, Daniel J. Gustafson (dgustaf@live.unc.edu) Ph.D. Candidate, Stephen E. Gent (gent@unc.edu) Associate Professor, Mark J.C. Crescenzi (crescenzi@unc.edu) Professor, Department of Political Science, University of North Carolina at Chapel Hill, 361 Hamilton Hall, Chapel Hill, NC 27599. An earlier version of this paper was presented at the 2017 Annual Meeting of the International Studies Association, February 2017, Baltimore, MD. The authors thank Cliff Morgan for valuable feedback on an earlier draft. They also thank Elizabeth Menninga, Johannes Karreth, Ryan Bakker, Santiago Olivella, Layna Mosley, and two anonymous reviewers for their helpful comments which greatly improved the the article.

Rwanda the following year. The Nairobi Agreement was supposed to end the Ugandan Civil War in 1985 but was never even implemented. The Lomé Peace Accord promised to end the Sierra Leone Civil War in 1999, but fighting continued until 2002. Almost every agreement signed by Afghanistan in the past three decades has been broken by one or more parties. Scholars and public officials deride these agreements and countless others as weak while praising long-lasting agreements such as the Good Friday Agreement as strong documents.

Yet, how much of the perception of civil peace agreements as weak or strong results from their observed duration? How could an agreement such as the Arusha Accords that was brokered as part of an extensive mediation process involving many third parties be so weak? Without being able to observe the counterfactual where Rwandan President Juvénal Habyarimana's plane was never shot down, we can know how much of the Accords' failure was due to his death rather than some inherent weakness in the agreement. This uncertainty suggests a need to measure the strength of an agreement separately from its duration.

There are several ways to measure the strength of a peace agreement, but each has its strengths and weaknesses. Given that even some strong peace agreements fail, the observed duration of an agreement is likely an imperfect indicator of its underlying strength. Specific characteristics of peace agreements give us some information about the strength or weakness of an agreement, but it is difficult to select a single characteristic that captures strength. An additive scale of provisions may be somewhat related to the strength of an agreement, but it weights all provisions equally. Treating all provisions the same is problematic because they likely do not all convey the same amount of information about agreement strength. Ceasefire provisions only result in a (potentially) temporary halt to the fighting, but power-sharing agreements require addressing underlying issues.

Given these issues, we take a new approach by treating agreement strength as a latent variable. Using Bayesian item response theory (IRT), we model the specific provisions within peace agreements as a function of an underlying latent agreement strength. We illustrate our new measurement strategy with an example of how scholars can apply it to substantive research questions by focusing on the question of whether external forces can influence peace agreement strength. The policy implications are clear: if outside actors can insert themselves and improve the strength of peace agreements, the chances of peace may improve. Alternatively, if external actors coerce belligerents to hastily sign agreements, the resulting document may fail to prevent future conflict.

To explore the underlying concept and determinants of peace agreement strength, this paper proceeds in three parts. First, we define peace agreement strength as a latent variable and discuss alternative measurements. Second, we present and explain our measurement strategy. Third, we present an illustrative analysis of the determinants of peace agreements focusing on the effects of third parties to validate our measure. We close by discussing the implications of our results for the study of conflict resolution more broadly.

Measuring Peace Agreement Strength

To best explain complex phenomena such as civil conflict termination and recurrence, we must first have a solid understanding of the qualities of negotiated settlements. While scholars have previously attempted to quantify the strength of peace agreements, we believe that analyses could benefit from an innovative measurement strategy. Before we introduce our measurement model, however, we first consider how negotiated settlements come about and discuss peace agreement strength as a theoretical concept.

At the most basic level, peace agreements in civil war settings seek to end a conflict

between a government and one or more nonstate actors. During negotiations, belligerents attempt to secure the greatest benefits for themselves while mitigating costs. Both parties have strong incentives to reach a settlement that halts the conflict because fighting inflicts great material costs. However, they may disagree about the specifics of an agreement. The government may balk at some rebel demands, and likewise, the rebels may find some of the government's preferred provisions unacceptable. Negotiations, which may include third party mediators, attempt to craft an agreement that leads to peace and that both parties will sign. Therefore, the peace process seeks to find a mutually agreeable settlement that produces the highest likelihood of sustained peace.

We define peace agreement strength as the degree to which a negotiated settlement addresses parties' potential grievances by encoding specific provisions. This is similar to the way in which Fortna (2003) defines agreement strength for international ceasefires. A strong agreement would address each of the potential causes of conflict, while a weak agreement would not. For rebel groups, fundamental grievances could stem from a desire for legal protections, political inclusion, or territorial autonomy. Governments generally seek a cessation of hostilities and disarmament by the rebels. A perfect agreement would address each of these concerns, while the worst possible agreement would solve none of these incompatibilities. Clearly, however, there are a range of possibilities between the best and worst potential agreements. We use the observable provisions within peace agreements to place them along this latent spectrum.

Consider, for example, the Arusha Accords signed in the summer of 1993 to end the three-year Rwandan Civil War. The talks were organized by the United States, France, and the Organisation of African Unity, and the resulting agreement contained several provisions considered important by existing literature on civil peace agreements. The Arusha Accords included provisions concerning the rule of law, repatriation of refugees,

and the integration of rebels into the national army. The rebel group, the Rwandan Patriotic Front (RPF), was granted participation in the Rwandan legislature and was given an equal number of cabinet posts as the former ruling party. While the agreement laid the groundwork for peace in Rwanda, it ultimately failed to prevent conflict recurrence, due in large part to the assassination of Rwandan President Juvénal Habyarimana. The eventual failure of the Arusha Accords shows that even agreements that are carefully crafted by well resourced stakeholders can fail. The disconnect between the amount of effort that went into reaching the Arusha Accords and their quick failure suggests that we cannot judge the strength of a peace agreement solely by observing its duration.

To assess the quality of peace agreements, researchers have largely conducted statistical analyses with the duration of the agreement as the outcome variable. While duration is certainly an outcome of interest for scholars, there is not a one-to-one mapping of agreement strength to duration. The durability of any given peace agreement depends upon factors beyond the scope of the agreement itself. For example, fluctuation in the global economy might induce conflict regardless of a given settlement's strength. Additionally, the death of Habyarimana appears to have played a role in the ultimate failure of the Arusha Accords to prevent conflict. While peace agreement strength and duration are certainly correlated and sometimes conflated, they are distinct outcomes. Our conception of strength reflects the inclusiveness of an agreement and is not necessarily related to an agreement's expected or actual duration.

Scholars have taken several approaches to examining the quality of peace agreements. Some have focused on the role of individual provisions. Previous work on peace agreement survival finds several provisions such as power-sharing arrangements, the degree of agreement institutionalization, and the specificity of the actual document are positively related to agreement duration (Hartzell, Hoddie & Rothchild 2001, Hartzell

& Hoddie 2003, Werner & Yuen 2005). Reid (2017*a*) takes a more holistic approach by showing that the degree to which an agreement is "context specific" is positively correlated with more durable agreements. While these studies have been foundational for understanding the importance of specific types of provisions, they focus on duration as the outcome of interest and cannot speak to the broader concept of agreement strength. Fortna (2003) uses both subjective coding and an additive index of provisions to show a positive relationship between agreement strength and durability for international peace settlements. While her approaches represent attempts to systematically analyze agreement strength, they each suffer from potential biases. The subjective coding of peace agreements may be prone to researcher bias, and additive indices are inappropriate because they either treat indicators as equally important to the latent construct, or suffer from disputes over the subjective weighting of different indicators (Smith & Spaniel 2018). Finally, Badran (2014) measures the strength of civil peace agreements using both an additive index and composite index produced via factor analysis. The composite index is an improvement on other attempts to characterize peace agreement strength, but still suffers from weighting issues and fails to preserve the variability in the raw data (DiStefano, Zhu & Mindrila 2009). Our definition of peace agreement strength is based upon the completeness of the agreements themselves. While this is not the only way to define peace agreement strength, it is an improvement on current conceptions.

Agreement Strength as a Latent Variable

We introduce a new measurement strategy to push forward the study and measurement of peace agreement strength. To do so, we turn to item response theory (IRT), a measurement strategy developed by the psychometrics literature. IRT models produce

estimates of an underlying attribute, such as ability in a given academic subject or quality of life, as represented by a series of observable indicators, such as questions on an exam or responses on a survey of health outcomes (Rasch 1980). In political science, IRT models are frequently used to measure the ideal points of individual legislators (Clinton, Jackman & Rivers 2004), building on the earlier NOMINATE system (Poole & Rosenthal 1985, Poole & Rosenthal 1997). Bayesian IRT models have also been used to measure political knowledge of survey respondents (Jackman 2000), bureaucratic agency preferences (Clinton & Lewis 2008), and mass political beliefs (Hare, Armstrong, Bakker, Carroll & Poole 2015). In the study of international relations and conflict, they have been used to measure states' nuclear capabilities (Smith & Spaniel 2018), regime type (Treier & Jackman 2008), human rights practices (Schnakenberg & Fariss 2014), the depth of preferential trade agreements (Dür, Baccini & Elsig 2014), the transparency of different governments (Hollyer, Rosendorff & Vreeland 2014), the transparency of private sector financial data within states (Copelovitch, Gandrud & Hallerberg 2018), and the scope of military alliance commitments (Benson & Clinton 2016). Scholars have used measurement models to improve theoretical accuracy, inference, and prediction (Bakker, Hill & Moore 2016, Carroll & Kenkel \noop{3001}Forthcoming, Fariss 2014, Gray & Slapin 2012, Pemstein, Meserve & Melton 2010).

For our measurement strategy, we employ the UCDP Peace Agreement Dataset (Harbom, Högbladh & Wallensteen 2006), which contains data on agreements from 1975-2005. While the data contain agreements in both international and civil conflicts, we believe that the two types of negotiated settlements may systematically differ. Therefore, we focus specifically on civil conflicts. These data contain three types of agreements: full agreements, partial agreements, and peace processes. A full peace agreement involves at least two participants in a conflict deciding to settle the entire incompatibility. A partial

agreement results when at least two parties in a conflict decide to resolve part of the incompatibility. Finally, a peace process agreement is an understanding that at least two parties in a conflict will take steps to work towards a resolution. Figure 1 shows the correlation between each of the provisions in the dataset.[1]

Consider the prospects of treating peace agreement strength as the outcome variable in an analysis. One approach would be to estimate a multivariate regression model where each provision is an outcome variable. However, this would be problematic due to our small sample size of 111 agreements, with 27 different provisions in the data.[2] Further, Figure 1 indicates that there is surprisingly little bivariate correlation between these provisions, with no two provisions having a correlation greater than $\pm$ 0.62. This pattern suggests that that not all provisions are related to the same aspect of peace agreements. No agreement has more than 18 out of 27 provisions, so simply adding all provisions together may result in biased measurements due to combining different concepts. Additionally, an additive index may mischaracterize the strength of an agreement by treating all provisions as equally meaningful.

Therefore, we estimate peace agreement strength as a latent variable that is a function of the provisions within it. Peace agreements have numerous provisions such as power-sharing arrangements, integration of former combatants into the armed forces, and language recognition that can be viewed as observable indicators of an underlying agreement strength. Although Badran (2014) finds that there are several dimensions to peace agreement strength,

[1]Implementation denotes whether an "agreement provided for the establishment of a commission or committee to oversee implementation of the agreement" and peacekeeping indicates whether an "agreement provided for the deployment of a peace-keeping operation." Implementation thus does not reflect third party involvement, while peacekeeping may. We estimate a model without peacekeeping in the Online Appendix, and results are unchanged.

[2]We omit the provision "border demarcation" because it refers to international borders, and hence does not appear in our sample of intrastate conflicts.

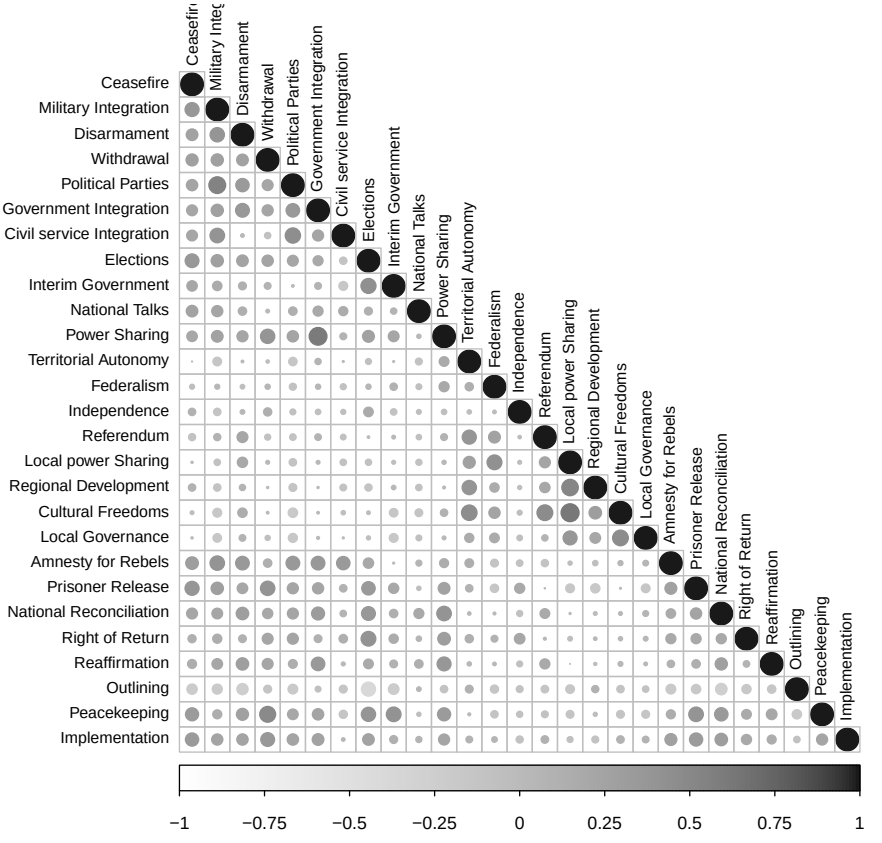

Figure 1. Correlation of all agreement provisions in the UCDP Peace Agreement Dataset (Harbom, Högbladh & Wallensteen 2006) for all agreements in our sample. Strength of correlation is represented by circle size and shade.

the peace agreement duration literature supports our decision to estimate a single latent measure of agreement strength.[3] Based on the argument that, *ceteris paribus*, stronger

[3]See the debate over the appropriate dimensionality for measuring US legislative ideology (Koford 1989, Poole, Rosenthal & Koford 1991) for an example of this question in other areas.

agreements should last longer (Fortna 2003), we argue that because these provisions are associated with longer lasting agreements, they can potentially be thought of as indicators for a one dimensional concept of agreement strength. Our model (which we discuss in more depth below) allows us to identify which indicators are positively related to our latent measure. Although we cannot be certain that our latent variable is capturing the strength of peace agreements, using indicators which are all positively correlated with agreement duration gives us confidence that we are indeed measuring agreement strength. Table 1 presents all provisions in the data and, where applicable, lists citations for their positive effect on agreement duration. This list represents the pool of candidate indicators for inclusion in our measurement model, but not all provisions are employed. We discuss this process at length in the section on agreement strength measurement.

We suspect that there is some latent underlying strength to peace agreements, and that this strength is expressed through the inclusion of these provisions. In other words, the stronger an agreement is, the more likely it is to have these provisions, which we refer to as indicators to be consistent with IRT literature. We estimate each indicator's relationship to the underlying dimension, which is the strength of a peace agreement in this case. For each indicator, we also estimate a discrimination parameter that determines how much the presence or absence of an indicator tells us about the agreement's underlying strength. For instance, 62.16% of agreements in our sample contain ceasefire provisions, while only 15.32% of agreements have provisions for the integration of former rebels into the civil service.[4] If both indicators are equally correlated with the latent strength of agreements, then the presence of civil service integration in a given agreement tells us more about its strength than the presence of ceasefire provisions does. Unlike the simple

[4]Full summary statistics for agreement provisions are available in the Online Appendix.

Provision	Citation
Ceasefire	
Integration of Rebels into Military	(Reid 2017*a*)
Disarmament	
Withdrawal of Foreign Forces	
Political Parties for Former Rebels	(Hartzell 1999)
Integration of Rebels into Government	(Hartzell 1999)
Integration of Rebels into Civil service	(Hartzell 1999)
Elections	
Integration of Rebels into Interim Government	
National Talks	
Power Sharing in Government	(Hartzell & Hoddie 2003)
Territorial Autonomy	(Hartzell 1999, Hartzell, Hoddie & Rothchild 2001)
Federalism	
Independence	
Referendum	
Local power Sharing	
Regional Development	(Hartzell 1999)
Cultural Freedoms	
Local Governance	
Amnesty for Rebels	
Prisoner Release	
National Reconciliation Efforts	
Right of Return for Refugees	
Reaffirm Earlier Agreement	
Outlining Peace Process	
Implementation of Peacekeeping	(Hartzell, Hoddie & Rothchild 2001, Fortna 2003)
Commission to Oversee Implementation	(Fortna 2003)

TABLE 1 *Peace agreement provisions in the UCDP peace agreements data, with citations for provisions that are associated with increased agreement duration. We omit border demarcation provisions from our analysis because no agreements in our sample of*

additive approach, the IRT model allows different indicators to contribute differentially to the strength of an agreement. Before we present results of our estimation, we briefly describe our initial application of the peace agreement strength measure: how external actors influence peace agreement strength.

Third Parties and Peace Agreement Strength

To what extent can third party actors shape the strength of peace agreements? We explore this question as a first-pass illustration of our measure of agreement strength. We consider four mechanisms by which external influences can affect the strength of a peace agreement. The first two, economic sanctions and threats of foreign aid revocation, can be thought of as indirect mechanisms that are sometimes used in cases of manipulative mediation or directive mediation (Beardsley, Quinn, Biswas & Wilkenfeld 2006, Touval & Zartman 1985). The second two, mediation and military intervention, are more direct ways for outside parties to become involved in conflict management.

We argue that states that are subject to economic sanctions are more likely to sign weak agreements. In a civil conflict setting, sanctions primarily affect the government because rebel groups are already reliant on smuggling and informal economic transactions, which third parties usually cannot disrupt. Economic coercion through sanctions shifts the incentives of the government, encouraging them to produce and sign agreements that they otherwise would not. An external state may threaten or impose sanctions to encourage the target state to produce a produce a peaceful settlement. Thus, sanctions raise the cost of non-agreement, making the government more willing to agree to any agreement with the rebel group.

Given the punishing costs that sanctions can generate, governments may have an

incentive to sign an agreement just to get relief from the sanctions. Consequently, governments may not be focused on signing the 'best' peace agreements they can when under economic sanctions. After the United States threatened to impose economic sanctions (Anna 2015) and a UN arms embargo (Nichols 2015) on South Sudan unless they ended their civil war, President Salva Kiir signed a peace treaty despite "serious reservations" (Dumo 2015). Kiir's concerns illustrate that he was aware of the dangers of the agreement, even going so far as to warn that "a poor agreement could backfire on the region." Crafting a strong peace agreement is a long and contentious process that involves bringing together all relevant stakeholders and attempting to reach a compromise that satisfies many different parties (Fortna 2003). Sanctioning states may underestimate the complexity of the situation and push for a faster resolution, leading to a weaker agreement.

An external state that desires a foreign civil war to end in the short-term may turn to economic coercion to encourage the belligerent state to sign a peace agreement. By imposing or threatening to impose sanctions, the external state can shift the belligerent state's incentives, making the costs of not signing an agreement greater than the costs associated with reaching a deal. However, because the warring parties do not organically reach this agreement, it may be drafted and signed in haste.

Sanctions work by cutting off access to international trade and other financial flows, but outside actors can also restrict government finances by suspending foreign aid payments. States that are dependent on this aid will be particularly receptive to these threats. Foreign aid is often allocated strategically, with countries receiving increased aid for democratizing (Alesina & Dollar 2000) or higher numbers of World Bank projects during their term on the UN security council (Dreher, Sturm & Vreeland 2009). The use of aid as a bargaining tool is in line with this behavior. Unfortunately, we cannot systematically observe threats to revoke aid the way we can with sanctions. Instead, we must settle for the degree to

which a state is dependent on foreign aid. While imperfect, this measure captures the ability of third parties to lean on governments to sign peace agreements in civil wars. Thus, peace agreements signed in states that highly depend on foreign aid will be weaker than agreements signed in other states.

While economic coercion through sanctions and foreign aid revocation should lead a peace agreement to be weaker on average, the relationship between mediation and agreement strength is more nuanced. In theory, mediation efforts should allow allow warring parties to come together and have structured conversations in an attempt to uncover and solve each belligerent's grievance. Here, the warring parties and mediation team might be able to structure a peace agreement that directly addresses areas of concern. However, in reality, mediation may actually serve as a substitution for full resolution (Werner & Yuen 2005). Additionally, mediation often leaves dyads worse-off in the long-term because of the artificial incentives that it imposes (Beardsley 2008). Because of this, we have reason to expect that mediation will produce weaker agreements on average.

Mediation will only be effective in generating strong agreements when mediators and belligerents work in an environment of trust and have strong incentives to contribute to the peace process. These conditions are satisfied when regional organizations serve as mediating parties. The states that make up regional organizations contain individuals that are likely to share important political and cultural characteristics with the belligerents. The similar identities can increase the actors' trust during negotiations (Olson & Pearson 2002, Wehr & Lederach 1991). Additionally, states that are close in proximity have strong incentives to prevent the spread of conflict (Kadera 1998). Because regional organizations as mediators facilitate trust and have material incentives to mitigate the likelihood of conflict recurrence, peace agreements signed in their presence will be stronger on average.

This logic relies on the process effects portion of Gartner's (2011) argument on

regional organization mediation and peace agreement duration. He notes that although regional mediators should be more effective in expectation, they disproportionately mediate intractable conflicts. This selection effect muddies any effect that regional organizations may have on peace duration, and accounting for the selection bias uncovers the true relationship. It is theoretically unclear whether or not the difficulty of settling a conflict is systematically related to a peace agreement's strength. It may be the case that more intractable conflicts tend to produce strong agreements. However, deep incompatibilities could disrupt the peace process, producing weaker agreements on average. Future work can focus on whether selection bias exists that can confound the relationship between regional organization mediation and peace agreement strength.

Intervention into an ongoing conflict can drastically increase its duration (Regan 2002) by introducing new veto players with different preferences than the primary combatants (Cunningham 2006). This effect may also lower the quality of any negotiated settlements reached in the conflict through two possible pathways. First, any agreement reached has to also satisfy the demands of external states in addition to those of the domestic combatants. This could result in weaker agreements that do not address the incompatibility between the initial combatants. Second, interveners who wish to extricate themselves from the conflict may push combatants to sign agreements, allowing them to withdraw. These agreements may be weaker than those signed more organically in conflicts without an internationalized dimension.

Model

Now that we have defined peace agreement strength as a latent variable and identified a substantive research question of interest, we present our measurement model of peace

agreement strength. First, we describe the model used to obtain estimates of latent peace agreement strengths. Second, we assess the plausibility of the measure. Finally, before presenting results of the full probability model, we introduce the variables used in our analysis on the effect of international forces on agreement strength.

Ultimately, we want to use our estimates of peace agreement strength to understand why some agreements are weak and others are strong. As estimates, these values of agreement strength are uncertain, and we must account for the uncertainty in our analysis.[5] In the best case scenario where this error is truly random, ignoring it will not bias coefficients but will bias standard errors downward. If it is not random, then ignoring it can bias both coefficient estimates and standard errors. Our approach accounts for both possibilities by estimating what Armstrong, Bakker, Carroll, Hare, Poole & Rosenthal (2014, 277-295) call a "full probability model,"[6] which allows the observed indicators for each agreement to determine the measured strength of the agreement while also letting the conflict-level explanatory variables explain variation in this strength across agreements. By including explanations for agreement strength in the model, we are able to share information across observations. Intuitively, two agreements signed at the end of territorial conflicts should be more similar than an agreement signed at the end of a territorial conflict and one signed

[5]The conventional procedure in this situation is to estimate two separate models: a measurement model to capture the latent construct and a regression model to explain variation in it. Unfortunately, this method ignores the uncertainty in the latent estimates. One way to overcome this limitation is to draw multiple samples from the posterior distribution of a latent construct and use them as the response variable instead of just the point estimate. For example, Fariss (2014) includes both the posterior mean and standard deviation of his latent human rights respect score so that users of the data can carry out this process.

[6]This term is different from "hierarchical IRT model," which can either mean a model with multiple latent constructs drawn from a higher level latent construct (Sheng & Wikle 2008) or simply a model with hyperpriors on the difficulty and discrimination parameters (Janssen, Tuerlinckx, Meulders & de Boeck 2000).

at the end of a governmental conflict. Estimating a full probability model lets us include the type of conflict an agreement was signed in, allowing us to incorporate this information into our estimates of agreement strength.

This model takes uncertainty around the estimated latent agreement strengths into account when estimating the effect of sanctions on agreement strength. This leads to a more conservative analysis because the explanatory variables have to explain variation in a range of agreement strengths instead of just a single value. This leads to more uncertainty in our estimates, so a strong effect for our explanatory variables should be interpreted as compelling support for our hypotheses.

Our full probability model is presented in Equations 1a-1i, where i indexes agreements, j indexes provisions, and k indexes conflicts. The observed indicators $\mathbf{X}$ are a function of latent agreement strength $\boldsymbol{\theta}$, multiplied by the discrimination parameters $\boldsymbol{\gamma}$, minus the difficulty parameter $\boldsymbol{\alpha}$. The discrimination parameter describes how much the presence of a given provisions tells us about the strength of an agreement, and the difficulty parameter tells us how strong an agreement must be to have a given parameter. Our explanatory variables $\mathbf{Z}$ enter into the model as hierarchical predictors on the mean of each agreement's strength, $\boldsymbol{\theta}$, with regression coefficients $\boldsymbol{\beta}$. In addition to these explanatory variables, the mean of $\boldsymbol{\theta}$ also includes a random intercept $\boldsymbol{\delta}$ by conflict, to account for a lack of independence between multiple agreements signed in the same conflict.[7] The means of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ have normal priors with diffuse normal hyperpriors, and the standard deviations of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, and $\boldsymbol{\theta}$ have diffuse half Cauchy hyperpriors. This choice of priors reflects our lack of theoretically driven expectations for the effect of our predictors. The regression

[7]The median conflict in the UCDP peace agreements data has 2 agreements, which means that ignoring this structure in our data would bias our estimates.

coefficients $\boldsymbol{\beta}$ have diffuse Student T priors.[8]

$$
\begin{aligned}
x_{ij} &\sim \text{Bernoulli}(\gamma_j \theta_i - \alpha_j) && \text{(1a)}\\
\theta_i &\sim \mathcal{N}(\delta_k + \mathbf{z}_i \boldsymbol{\beta}, \sigma_\theta) && \text{(1b)}\\
\boldsymbol{\alpha} &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha) && \text{(1c)}\\
\boldsymbol{\gamma} &\sim \mathcal{N}(\mu_\gamma, \sigma_\gamma) && \text{(1d)}\\
\boldsymbol{\delta} &\sim \mathcal{N}(\mu_\delta, \sigma_\delta) && \text{(1e)}\\
\mu_\alpha, \mu_\gamma &\sim \mathcal{N}(0, 25) && \text{(1f)}\\
\mu_\delta &\sim \mathcal{N}(0, 5) && \text{(1g)}\\
\sigma_\alpha, \sigma_\gamma, \sigma_\delta, \sigma_\theta &\sim \text{hCauchy}(0, 5) && \text{(1h)}\\
\boldsymbol{\beta} &\sim t(4, 0, 1) && \text{(1i)}
\end{aligned}
$$

The standard IRT model is unidentified due to possibility of infinite rotations which could fit the data, so we place two identification restrictions on the model (Bafumi, Gelman, Park & Kaplan 2005). First, the sign on the discrimination parameter $\boldsymbol{\gamma}$ is constrained to be positive. This is based on the assumption that all included indicators are coded so that their presence indicates a stronger agreement, while their absence denotes a weaker one. We first estimate our measurement model on *all* agreement provisions, and then evaluate how well this assumption fits our data. Second, to identify our model, we fix the values of $\boldsymbol{\theta}$ for two peace agreements: the DUP/SPLM Sudan Peace Agreement between the Democratic Unionist Party and the Sudanese People's Liberation Movement, and the Tripoli Agreement between the government of the Philippines and the Moro National

[8]The separate measurement and regression model approach, which we refer to as a standalone IRT model, splits Equations 1a and 1b and their respective priors into two distinct models run sequentially. We estimate this model and present results in the Online Appendix, finding that coefficient estimates are substantively similar but with smaller credible intervals because this specification ignores the uncertainty in our latent estimates.

Liberation Front.

Setting the value of θ for these two agreements anchors the latent construct and ensures that our results are in the 'correct' orientation, with stronger agreements above zero, and weaker ones below. We set $\theta = -1$ for the DUP/SPLM Sudan Peace Agreement because it has only 2 provisions, and we set $\theta = 1$ for the Tripoli Agreement because it has 12 provisions. The most provisions that any agreement has is 18, and because our model defines stronger agreements as those with more provisions, the DUP/SPLM Sudan Peace Agreement can serve as a 'weak' anchor, while the Tripoli Agreement is a 'strong' one. Picking agreements based on the number of provisions they have may raise the question of why we are not simply using the number of provisions to measure agreement strength. As discussed earlier, counting the number of provisions assumes all provisions contribute equally to agreement strength. This assumption can be particularly problematic for provisions which are present in almost no agreements or nearly all. Fixing the value of θ for these two agreements merely orients our latent scale; it does not determine the strength of these two agreements.

Agreement Strength Measurement

Before presenting the estimates produced by our model, we pause to assess the validity of our measurement strategy. We first estimate a model that includes all provisions as indicators in the measurement model, with their presence coded 1 and their absence coded 0. The identification restriction that γ must be positive is based on the assumption that all indicators have a positive effect on the underlying quantity. To verify this, we plot the densities of the indicator discrimination parameters to ensure that this is a reasonable constraint (Bafumi et al. 2005, 178). If all indicators have a positive effect on the latent

quantity, then the densities of all parameters should be well to the right of zero. As the parameters are constrained to be positive, no densities will be to the left of zero, but an indicator that does not have a positive relationship with the latent quantity will have a density that bumps right against zero. Any indicator that is concentrated against zero may be representative of a different latent quantity than that represented by indicators with γ values greater than 0.

We see that the majority of the density for territorial autonomy, federalism, independence, referendum, local power sharing, regional development, cultural freedoms, and outlining the peace process is concentrated right at zero. As there are several indicators here, this suggests that they may be indicators of some underlying dimension other than agreement strength. With the exception of peace process outlining, all of these provisions are related to local autonomy in some way or another. This suggests that there is a second latent dimension connected to territoriality and self-determination. While this suggests opportunities for future research, we focus on the agreement strength dimension.

We subsequently remove such 'territorial' provisions with $\gamma \approx 0$ from our analysis and do not present results from this specification as the measurement of the latent concept would be biased. Given that the indicators included in the final model are strongly associated with increased agreement duration in the literature, we can be more confident that this latent dimension reflects the underlying strength of peace agreements.

We next present the measurement model's difficulty and discrimination parameters, α and γ, from the full probability model estimated using only relevant provisions. Examining these parameters helps us to understand what each provision tells us about the latent strength of a peace agreement. This is an important exercise because there is no simple test to check whether the latent construct that we have created actually aligns with our concept of peace agreement strength. Instead, we need to see whether the parameters in

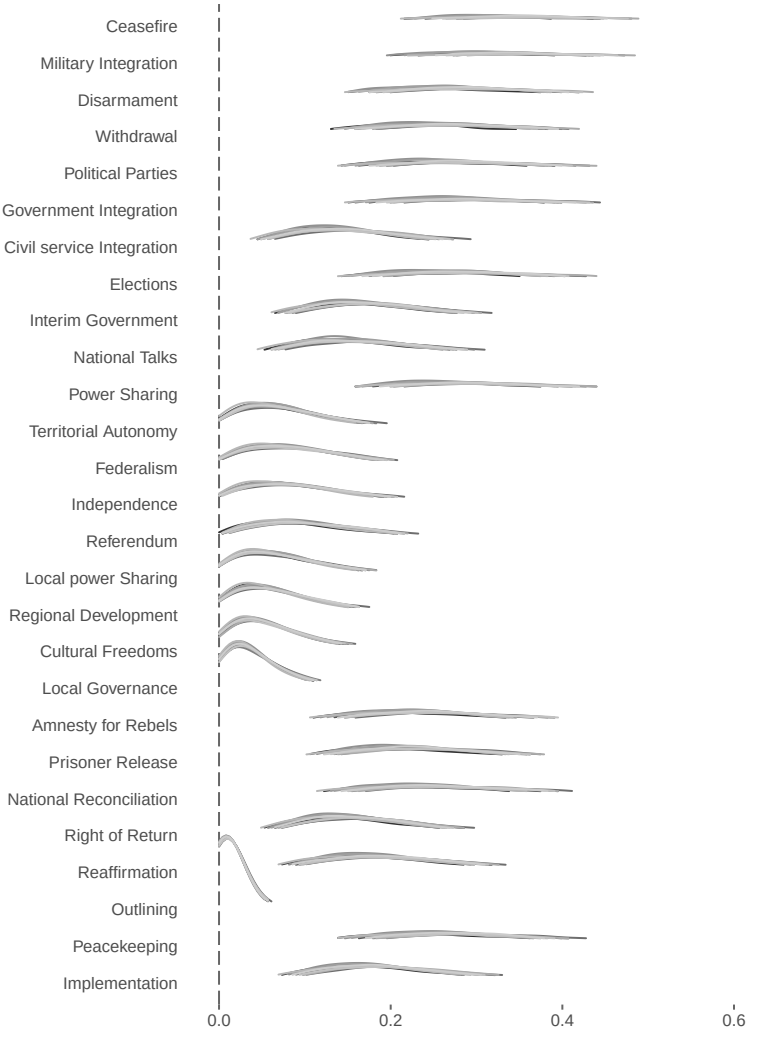

Figure 2. Discrimination parameters for all agreement provisions. Parameters with the majority of their density near zero represent a different latent dimension. We remove these parameters from our analysis

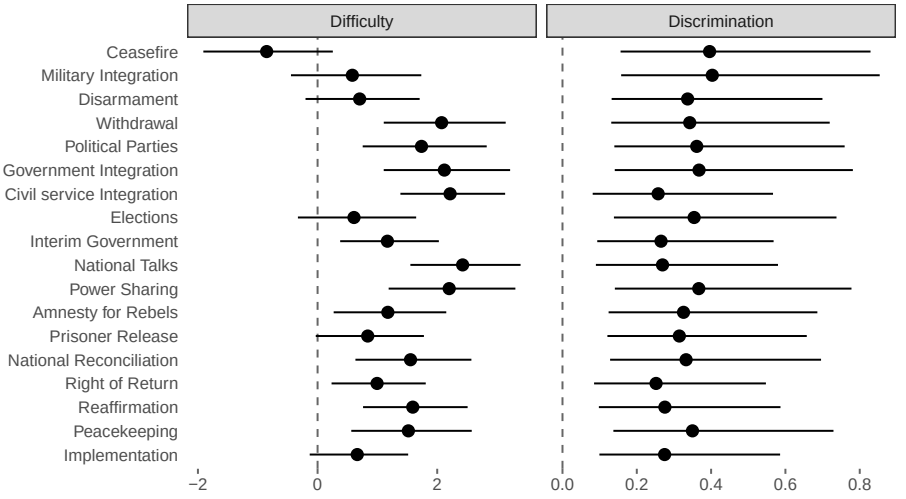

Figure 3. Difficulty (α) and discrimination (γ) parameters in the measurement model. The difficulty parameter controls the location of the item characteristic curve's inflection point, while the discrimination parameter controls the slope.

the model align with our theoretical expectations of how observed indicators should relate to strong and weak agreements.

Figure 3 presents the posterior means of the difficulty and discrimination parameters in the measurement model. The higher the value of the difficulty parameter, the higher the baseline level of agreement strength required for a provision to be present. This means that an agreement has to be very strong for power sharing or civil service integration provisions to be included, but even a very weak agreement is likely to have ceasefire provisions due to its low parameter estimate of -0.85. As seen in the Online Appendix, ceasefire arrangements are the most common provisions, appearing in 62.16% of agreements. Given their prevalence, it makes sense that agreements do not have to be very strong to include ceasefire provisions.

The higher the value of the discrimination parameter, the steeper the item characteristic

curve (ICC) for that provision. Steeper ICCs indicate provisions that are better discriminators between strong and weak agreements. The provision with the highest discrimination is military integration with a parameter estimate of 0.4. This means that military integration are the best provision for discriminating between weak and strong agreements. Given their frequency, ceasefire provisions are a surprisingly good discriminator, with an estimate of 0.4. Taken together, these two parameter estimates mean that any agreements without ceasefire provisions must be exceptionally weak.

Power sharing agreements have a very high difficulty parameter value of 2.2 and a relatively high discrimination parameter value of 0.37. This means that agreements must be strong to have power sharing provisions, and that the presence or absence of power sharing provisions tells us much about the strength of a given agreement. These parameter values align with the findings from the peace agreement duration literature that power sharing agreements have a significant positive impact on the duration of peace (Hartzell & Hoddie 2003).

The relationship between the provisions included in peace agreements and their underlying strength revealed by these parameters largely aligns with our expectations. As such, we can be confident that our latent construct really does reflect what we would analytically describe as the strength of a peace agreement. We now move on to our findings of how international attention and involvement can affect these underlying peace agreement strengths.

As an initial face validity check, we note that the correlation between our latent measure of agreement strength and the comprehensiveness of an agreement is 0.42. This measure of comprehensiveness comes from the peace agreements data (Harbom, Högbladh & Wallensteen 2006) and is a three point ordinal variable denoting whether an agreement is

a process, partial, or full agreement, with more comprehensive agreements coded higher.[9] More comprehensive agreements should address more of the underlying differences behind a conflict and should be stronger as a result. This positive correlation suggests that our latent construct is properly oriented so that higher values represent stronger agreements.

The correlation between our latent strength measure and a simple additive index of provisions is 0.89. While stronger than the correlation with the comprehensiveness measure, this correlation is still not perfect. Mathematically, differences between the two can be explained by the varied difficult and discrimination parameters in the measurement model. Substantively, this means that the nuance introduced by a measurement model tells us more about the underlying strength of a given agreement because it accounts for the fact that not all provisions are equally representative of strength. Just as an additive index is contains more relevant information than a three point ordinal variable, our latent strength measure is a similar improvement.

Figure 4 presents estimates for all agreements in our sample, along with associated measures of uncertainty.[10] The two point estimates with no uncertainty are the two agreements whose strength we fix to identify and orient our model.[11] Interestingly, two frequently discussed agreements in the literature have opposite positions from what we would expect. The Arusha Accords are often held up as an example of a weak agreement that failed, leading to the resumption of hostilities and large-scale civilian killings. However, in our scale, they are one of the strongest peace agreements. The Good

[9]See the Online Appendix for the full description of each type of agreement.

[10]The agreement strength values presented here are averages of the results from all five imputed datasets.

[11]Agreement strength estimates from the standalone IRT model are presented in the Online Appendix. The extra information included in the full probability model produces a much larger range of agreement strength values, allowing for more meaningful inference on the effect of international involvement on agreement strength.

Friday Agreement, which ended the Troubles in Northern Ireland, is frequently considered to be a strong agreement responsible for the long-lasting peace. Yet it is in the lower half of the spectrum. We return to this puzzling finding further in our discussion.

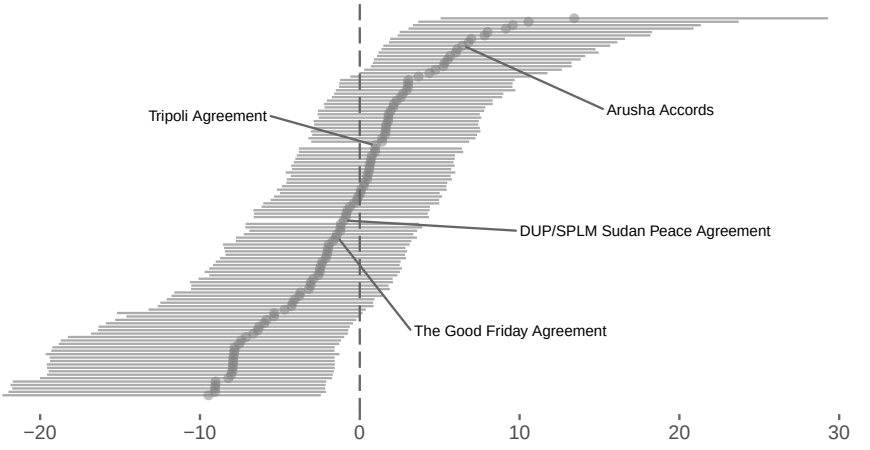

Figure 4. The posterior mean of latent agreement strength is represented by the points, while the lines denote 95% credible intervals. The observations without any uncertainty are the DUP/SPLM Sudan Peace Agreement and the Good Friday Agreement, whose values are fixed and thus not estimated.

The distribution of agreements along this latent scale is relatively invariant to different choices of agreements for the strong and weak identification restriction. In addition to the results from the model above, Figure 5 presents the distribution of agreement strength for models with three different sets of identification restrictions. The position of the points in Figure 5 represents how far an agreement has moved in the order compared to the model in Figure 4. The *x* axis is the order in the main model whose results are presented here, while the *y* axis represents the order in models with three different choices of weak and strong agreements to identify the scale. Points exactly on the diagonal indicate an agreement whose position in the ranking of agreements is identical under both sets of identification

restrictions.

An unstable measure would see few points near the diagonal, while a stable one would see many points along the diagonal. Most of the points in Figure 5 are relatively close to the diagonal. This suggests the latent scale of agreement strength is not sensitive to choice of identification restriction. We present the coefficients for agreement strength predictors from each of these models in the Online Appendix; the magnitude and direction of estimates is stable with regard to choice of identification restriction.

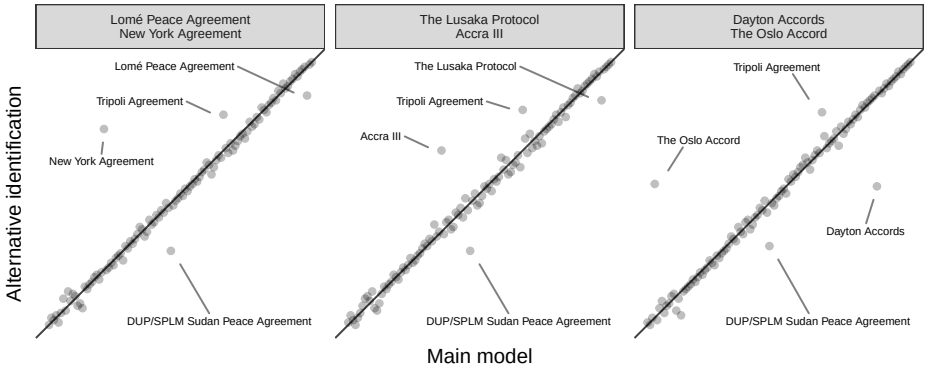

Figure 5. Comparison of agreement strength estimates from models with three different sets of end points selected as identification restrictions. Points below the diagonal indicate agreements that are ranked higher in our main model, while those above the diagonal indicate agreements ranked higher under an alternative identification strategy. Agreements which shift more than 10 places in the ranking are labeled.

Looking at Figure 5, there are relatively few agreements which move more than 10 places in rankings between our main model and ones with alternative identifying agreements. All of the agreements which move more than 10 places in rank are identifying

agreements in one of our models. All other agreements do not move more than 10 places in the ranking, and some do not move at all between identification strategies. This pattern suggests that the ranking of the agreements chosen as identification restrictions may vary significantly, but that the ordering of the remainder of agreements will not. Importantly, the Good Friday Agreement and the Arusha Accords do not deviate more than 1 place in rank across any of the models. The results for these oft-studied agreements are stable regardless of identification restriction, meaning that we can measure and study their *strength* directly, regardless of their eventual duration. Consequently, it may be prudent to not select theoretically interesting or especially policy relevant agreements for identification restrictions.

Comparing estimates of agreement strength from the full probability model and the additive index demonstrates the benefits of our approach. We specifically look at the two agreements with the most provisions, and hence, the highest estimated strength with an additive index. The Global Ceasefire agreement between Transitional Government and the Forces pour la defence de la democratie (CNDD-FDD) of Mr. Nkúrunziza in Burundi and the Sudan Comprehensive Peace Agreement in Sudan both have 18 of the total 27 provisions in the data. An additive index approach would lead us to conclude that these two agreements have similar, if not identical, strengths. In fact, the full probability model estimates that the former has a strength of 13.42 while the latter has one of 5.31. The two agreements share only 13 of 18 provisions, and the provisions present in the former include national talks and civil service integration, which are two of the provisions with the highest discrimination parameter estimates, so their absence from the Sudan Comprehensive Peace Agreement indicates that it is weaker.

These two agreements demonstrate the nuance introduced by our measurement model compared to a simple additive index. Agreements can have the same number of

provisions, but if they have different provisions, their strengths may radically vary as in the above example. Our measurement model thus allows us to identify differences between agreements with the same number of provisions, which we cannot do with an additive index. Even when two agreements share the same provisions, our full probability model offers advantages over a traditional IRT model. While the agreements may share the same provisions, they will have different predictor values, and so they will receive different strength values.

Explanatory Variables

To capture whether a state faces economic sanctions in connection with an ongoing conflict, we rely on the Threat and Imposition of Economic Sanctions (TIES) dataset (Morgan, Bapat & Krustev 2009, Morgan, Bapat & Kobayashi 2014). The TIES data contain not only records of sanctions that are actually implemented, but also the threat of sanctions in cases where sanctions themselves are never actually enacted. The mere threat of sanctions is enough the shift the incentives of a state and make a deal more likely even if they are never actually implemented (Bapat & Kwon 2015). Many sanctions achieve their purpose without having to actually carry them out, so the data allow us to capture the threat of sanctions. The danger of loss of economic activity is often enough to elicit change, as in the South Sudanese case where the possibility of the loss of economic activity due to new sanctions was enough to convince the government to sign. We use the TIES data to create a categorical variable that takes on a value of 0 for *no sanction*, 1 for a *unilateral sanction* episode, and 2 for a *multilateral sanction* episode.

To measure whether an agreement was signed as part of a mediation process or not, we use the Civil Wars Mediation (CWM) dataset (DeRouen, Bercovitch & Pospieszna

2011, DeRouen & Bercovitch 2012). We code a categorical variable that takes on a value of 0 for *no mediation*, 1 for *mediation*, and 2 for *regional mediation*.[12] Any agreement that is coded as a 1 on this variable indicates that mediation occurred, but there was no regional organization that participated as a mediator.

To operationalize a state's dependence on foreign aid, we use the fraction of a state's GNI that comes from official development assistance (World Bank 2018) to construct the variable *aid*. This variable is measured relative to the national economy because absolute amounts do not tell the complete story. A state with a small economy would suffer greatly from the loss of even aid revenues, while an economically powerful state could absorb the loss of much more money before feeling any pain.

To determine whether a conflict is subject to third party military intervention, we use the International Military Intervention (IMI) Dataset (Pickering & Kisangani 2009). These data define military intervention as the "purposeful . . . result of conscious decisions by national leaders," which means that an intervention is evidence of foreign interest in a civil war and not merely the result of accidental entanglement.[13] We use the IMI to construct the dummy variable *intervention*, which denotes whether foreign military forces were engaged in an intervention in the country on the date an agreement was signed.

[12]Note that this means we potentially over-count mediation as the available data do not allow us to match on conflict, just state. We perform median imputation on mediation episodes which are missing start or end dates. See the Online Appendix for details.

[13]As a robustness check, we estimate a model using the *internationalized internal armed conflict* variable from the ACD. The effect of intervention changes due to the less precise nature of this variable, but all other results are unaffected.

Control Variables

In addition to our explanatory variables, we employ a number of control variables to account for other sources of variation in peace agreement strength. We are interested in the ability of external efforts by the international community to improve peace agreement strength, so we want to make sure that we are controlling for domestic influences that may be responsible for some of the variation in agreement strength.

To control for the possibility that stronger states will sign stronger agreements, we use relative political reach (*RPR*), from the Relative Political Capacity data (Arbetman & Kugler 1997, Kugler & Tammen 2012). Although GDP is commonly used to measure state capacity, RPR is better suited to our analysis because it is meant to properly measure cases where resource rich states have less administrative capacity than less wealthy counterparts. RPR is a more nuanced measure of state capacity than and relates to the idea that more capable states should have higher expected compliance from their populations, producing stronger agreements.

RPR represents the ability of a government to "mobilize human resources" and "the degree to which the population accepts the presence of government in their lives (Kugler & Tammen 2012, 20-24). Monitoring and enforcing a peace agreement requires extensive cooperation from a country's population; the state is reliant on the willingness of people to report the stirrings of a new insurgency. When the government is unable to convince its people to support its efforts it will face great difficulty in detecting and stopping any resurgent rebel movements. RPR thus serves as an important control for how strong a peace agreement signed by a given government will be.[14]

[14]GDP is one of the components of RPR, so we do not include it as a control variable in our analysis.

As there is evidence that the characteristics of a conflict can influence the survival of peace agreements (Fortna 2003, Werner & Yuen 2005), we want to control for the possibility that they can also affect the strength of agreements signed by belligerents. We include a measure of the underlying incompatibility within a conflict, treated as a dummy variable that takes on a value of 0 for *territorial* conflicts and a value of 1 for conflicts over the *government*. Because mediation efforts often target the most intractable cases (Greig 2005, Gartner & Bercovitch 2006), we include whether the conflict's *cumulative intensity* has exceeded 1,000 battle-deaths at the time of the agreement's signing (Gleditsch, Wallensteen, Eriksson, Sollenberg & Strand 2002, Themnér & Wallensteen 2014) to control for this selection effect. Due to decreasing security commitments after the end of the Cold War, Western governments can more credibly threaten the withdrawal of foreign aid (Bearce & Tirone 2010), so we include a dummy variable measuring whether an agreement was signed *post cold war*. Following standard practice, we also measure the government's *polity2* score (Marshall, Gurr & Jaggers 2014) at the time of agreement signing.

Results

In this section we present and discuss results from our full probability model. We estimate bivariate models for each of our explanatory variables, a model with all four explanatory variables, and a model with all control variables. Our discussion of the latent agreement strength measure in the following section is based on results from this last model as it provides the most accurate estimates by controlling for alternative sources of variation.[15]

[15]The indicator *peacekeeping* denotes whether an "agreement provided for the deployment of a peace-keeping operation" (Harbom, Högbladh & Wallensteen 2006). As it does

We estimate our model using the Stan probabilistic programming language (Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, Brubaker, Guo, Li & Riddell 2017) in R (R Core Team 2018) via the RStan interface (Stan Development Team 2017).

Due to missingness in the explanatory variables, we multiply impute the missing values using the `mice` package (van Buuren & Groothuis-Oudshoorn 2011). We generate 5 imputed datasets, run two chains on each, and then perform inference on all 10 chains pooled together, averaging over the uncertainty in different imputed values (Little & Rubin 2002, 217-218).[16] We run each chain for 30,000 warmup iterations, followed by 60,000 sampling iterations; all results presented are from the sampling iterations. All continuous predictors are centered and scaled to aid with mixing.[17]

Agreement Strength Explanation

Now that we are satisfied that we are in fact capturing the latent strength of peace agreements, and doing so reliably, we turn to exploring the effect of international attention on agreement strength. In the remainder of this section we report and interpret our

not denote whether peacekeeping occurred after the signing of an agreement, it does not introduce post-treatment bias. However, it could indicate a level of third-party interest in the conflict, raising concerns that international factors are used both to measure and explain agreement strength. We estimate a model omitting this indicator. Results are substantively similar and presented in the Appendix.

[16]Although it is possible to employ a model that jointly specifies the probability of an observation's absence alongside the parameters of interest, doing so is unnecessary in this case. When the proportion of missing information in a dataset is low, this "uncongeniality" between separate imputation and analysis models does not affect inference of imputed data (Meng 1994). The proportion of missing data in our dataset is 0.01, so we are confident in the validity of our inferences.

[17]Standard diagnostics provide good evidence that our Markov chains have achieved convergence and explored the full parameter space of the posterior distribution. A full discussion and report of diagnostics is presented in the Online Appendix.

estimates of this effect.

We summarize the samples from the posterior distribution for our full probability model in Table 2. The design matrix $\mathbf{Z}$ in the regression model does not contain an intercept term, so we include the mean of the random intercept μ_δ in our results as the grand mean of the regression model. We present the posterior mean and 95% credible interval for each predictor.[18] Models 1-4 include our explanatory variables individually, while Model 5 includes all explanatory variables. Model 6 adds our control variables. The results are relatively stable across all specifications.

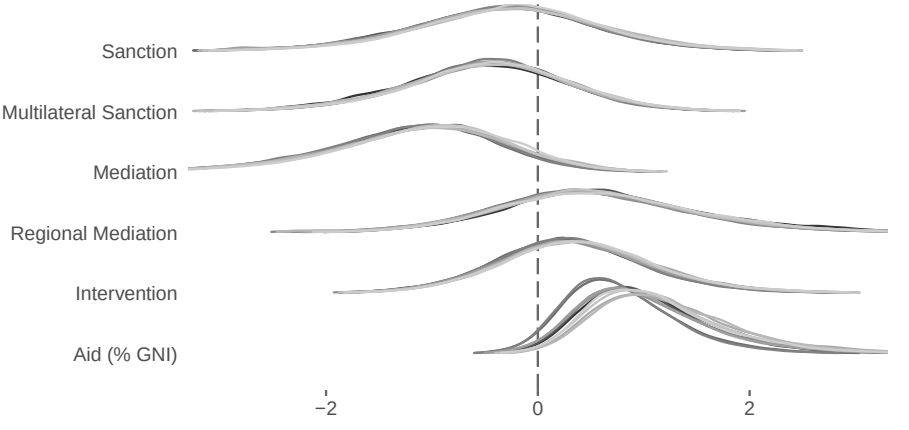

Figure 6. Posterior distributions for parameter estimates of our explanatory variables. Each shaded line represents a different chain, and the overlap between the lines indicates that the chains have converged to the stationary distribution. Although we display only our explanatory variables, these results are from Model 6 with control variables included.

As we do not have a parametric hypothesis test threshold to evaluate the significance of our results, we want to be able to assess the effect magnitude and direction for each

[18]We also calculate the 95% highest posterior density interval around these estimates in the Online Appendix. The results are substantively similar.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Sanction	−0.23				−0.27	−0.31
	[−1.79; 1.32]				[−1.98; 1.33]	[−2.30; 1.54]
Multilateral Sanction	−0.43				−0.48	−0.50
	[−1.92; 0.75]				[−2.20; 0.87]	[−2.27; 1.09]
Mediation		−0.87*			−0.93	−1.19
		[−2.22; −0.01]			[−2.66; 0.34]	[−3.36; 0.39]
Regional Mediation		0.60			0.55	0.58
		[−0.61; 2.10]			[−1.06; 2.50]	[−1.32; 2.85]
Intervention			0.49		0.43	0.37
			[−0.41; 1.53]		[−0.81; 1.95]	[−1.11; 2.04]
Aid (% GNI)				0.70*	0.97*	1.06*
				[0.18; 1.69]	[0.16; 2.35]	[0.02; 2.71]
Government						0.27
						[−1.45; 2.27]
Cumulative Intensity						0.91
						[−0.89; 3.18]
Post Cold War						0.62
						[−1.12; 2.66]
RPR						−0.64
						[−2.05; 0.52]
Polity2						−0.52
						[−1.88; 0.59]
μ_δ	0.02	0.42	−0.30	−0.04	0.00	−1.32
	[−2.90; 2.47]	[−1.33; 2.16]	[−1.69; 1.35]	[−1.86; 1.49]	[−2.94; 2.78]	[−5.49; 2.86]

* 0 outside 95% credible interval

TABLE 2 *Posterior density of parameter estimates for explanatory variables. The point estimates are posterior means and represent the most probable value for the relationship between each variable and agreement strength.*

predictor. To accomplish this, we present our results graphically in Figure 6.[19]

We find support for some of our expectations pertaining to the relationship between international third-party actions and civil peace agreement strength. The posterior means of both *sanctions* and *multilateral sanctions* are negative. The model suggests that the probability that *sanctions* are associated with weaker peace agreements relative to *no sanctions* is 64%, and the same probability for *multilateral sanctions* is 73%. However, given that the β distributions for these two variables have substantial density around 0, we are unable to state that economic sanctions or the threat thereof lead agreements to be weaker on average.

We find that *mediation* has a negative impact on agreement strength relative to *no mediation* with a probability of 92%, while *regional mediation* is positive relative to *no mediation* with a probability of 68%. This lends support for our expectations, although there is a large amount of uncertainty over the effect of *regional mediation* relative to *no mediation*. Importantly, however, there is a substantial difference between the posterior estimates for *mediation* and *regional mediation*, providing evidence that the latter is associated with stronger agreements than the former. The model indicates that the probability that *military intervention* is positive is 65%, contradicting our expectation. Finally, our model suggests that an increase in *foreign aid* dependence has a probability of 98% of being associated with stronger agreements. This finding does not conform with our theoretical expectations, but this may be due to measurement error because we are unable to directly measure the true concept of interest—threat of foreign aid revocation.

The units of scale for agreement strength are not inherently meaningful because they are the result of a latent variable estimation. Accordingly, they should be thought of relative

[19]A standard coefficient plot is included in the Online Appendix.

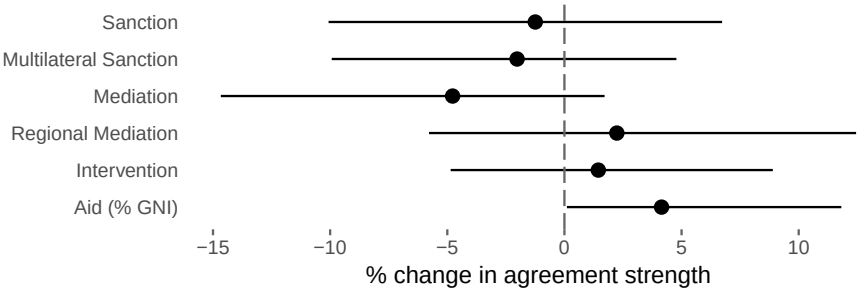

Figure 7. Estimates of the marginal effect of a one unit shift in predictors on agreement strength, with 95% credible intervals. A 5% change in agreement strength means that a dummy variable's presence moves the estimate 5% up the range of agreement strength, or that a one unit change in a continuous variable's value moves the estimate the same distance.

to the total extent of agreement strength values. Interpreting the relationship between peace agreement strength and both types of sanctions and mediation along with military intervention is relatively straightforward because they are dummy or factor variables, so their marginal effect is just the result of their presence or absence relative to the excluded category. The effect of aid is less straightforward, but relatively simple, because it is standardized to have mean 0 and unit variance. Figure 7 presents the median effect of a one unit shift in our predictors on an agreement's position within the range of agreement strength estimates. The median effect of an agreement being signed under the duress of economic sanctions is -0.28, which shifts an agreement's strength 1.24% downward in the distribution of agreement strengths, which is not as far down as the -0.46 median effect of multilateral sanctions which shifts agreements 2.02% downward. Foreign aid, the only predictor with more than 95% of its posterior distribution on the same side of zero, produces a change of 0.95 for each one unit shift in scaled aid, which translates to a 4.15% shift upward in the distribution of agreement strengths.

Although we find some support for our expectations, our null findings suggest important

implications. We are unable to find strong associations between the presence of economic sanctions or military intervention and the strength of an agreement. These null findings are surprising given the abundance of literature that shows how states are able to exercise some control over foreign countries' domestic politics by using these tools. Additionally, we find that regional mediation is associated with stronger agreements than non-regional mediation, but our model suggests that there is very little difference between the strengths of agreements associated with regional mediation and no mediation. The murky findings concerning the presence of third-party mediation and the null findings for economic sanctions and military intervention indicate the need for more research on the determinants of agreement strength. Future work may be benefit from carefully considering the types of domestic and foreign political actions that can alter the body of a peace agreement.

Discussion

Our analysis suggests a need for more consideration of the relationship between provisions, agreement strength, and the duration of peace agreements. We find that agreement strength—treated as a function of the provisions within the document—is negatively correlated with agreement duration ($\rho = -0.4$). It may be true that, *ceteris paribus*, stronger agreements last longer, but we are unable to control for *all* relevant factors. Accordingly, the variables we are omitting in our analyses of agreement duration may be sufficiently important to bias our estimates of what matters for agreement strength.

This unexpected negative correlation between agreement strength and duration raises an important question about the validity of our measure. Given that the majority of the literature argues that stronger agreements should last longer, one possibility is that we are not correctly measuring agreement strength. Another possibility, however, is that stronger

agreements present belligerents with more encoded constraints over their future behavior, and parties to the agreement are more likely to renege when there are multiple provisions. Additionally, measurement error may exist in the duration of peace agreements because the coding rules used to determine whether or not an agreement has ended are imprecise.

One common validation approach in this situation is to replicate existing studies using the new measure. For example, Smith & Spaniel (2018) use IRT to develop a latent measure of nuclear capability and replicate Jo & Gartzke (2007), which uses a simple additive index of nuclear technology provisions to explain which states develop nuclear weapons. Similarly, Carroll & Kenkel (\noop{3001}Forthcoming) employ machine learning methods to generate a new measure of expected interstate dispute outcome, and show that it improves predictive accuracy over the widely used CINC score (Singer, Bremer & Stuckey 1972, Singer 1988). This comparison approach is a straightforward way to demonstrate the utility of a new measurement strategy.

Unfortunately, we are unable to replicate a previous study because, to the best of the authors' knowledge, this study represents the first attempt to systematically measure the strength of peace agreements using more than one or two provisions. In related work, Hartzell (1999) codes the institutionalization of a peace agreement by determining whether it has rules regarding the use of coercive power, the distribution of political power, and the structuring of distributive policy. Hartzell & Hoddie (2003) use political, military, territorial, and economic power-sharing to code the institutionalization of an agreement. It is not clear how to translate these coding rules to the provisions in the UCDP Peace Agreements data, making comparison with these studies difficult. Fortna (2003) constructs a subjective measure of agreement strength, as well as an additive index of agreement provisions, but her sample is of interstate conflicts, so we cannot make comparisons to our measure of intrastate conflict agreement strength. Instead, our contribution lies in opening

up new avenues of research into conflict resolution, which we discuss in our conclusion.

Although we are unable to replicate previous work, we believe that the surprising latent strength values of some agreements in our sample offer insight into how we study conflict resolution. The position of the Arusha Accords near the top of our scale and the Good Friday Agreement below the middle are particularly curious. The model implies that the Arusha Accords is strong partially because several of its provisions have very high discrimination parameters in our measurement model such as military integration, peacekeeping operations, and elections. Thus, the agreement encoded a number of theoretically peace-improving provisions despite its short existence. Our model suggests that the Good Friday Agreement is not strong despite its persistence for two reasons. First, the Good Friday Agreement does not contain a provision for a ceasefire, which has a low difficulty parameter and a high discrimination parameter. Our model punishes the Good Friday Agreement for lacking a ceasefire provision given that, in our sample, ceasefire provisions are relatively easy to come by and tell us quite a bit about the latent strength of an agreement Additionally, several of the provisions in the Good Friday Agreement such as those pertaining to cultural freedoms and a referendum were dropped because we deemed them to relate to a different latent dimension. This second dimension of peace agreements presents one interesting avenue for future research.

Conclusion

By employing Bayesian IRT, we are able to measure and explain the strength of peace agreements. This novel approach allows us to build upon earlier work by Fortna (2003) without having to rely on simple additive indexes or subjective codings of agreement strength. Our measure of agreement strength contains substantial variation, even among

agreements with the same number of provisions, indicating that it is better at capturing qualitative differences between agreements. In contrast to subjectively weighting specific provisions, a Bayesian full probability model of agreement strength offers a principled way to exclude irrelevant provisions while allowing the data to determine the relationship between individual provisions and agreement strength. With this measure, we can now systematically investigate the dynamics of peace agreement strength. Doing so reveals some unexpected relationships between the international community and agreement strengths.

We believe that our measurement strategy improves upon current operationalizations of peace agreement strength, but the decision about which measure to use is fundamentally dependent on the research question at hand. Our measure is essentially a consolidation of the information present in peace agreements that pertain to a single dimension of peace agreement strength. Because of this, this measure is useful when research questions focus on the strength of negotiated settlements as a concept. Our measure is not well suited for research questions that are concerned with the causes or effects of individual provisions present in peace agreements. If a theory is concerned with specific policies such as power-sharing agreements, this measure is not appropriate. Additionally, our measure of peace agreement strength cannot speak to issues of implementation or observation outside of what is contained in the provisions. Our measure includes information about agreed-to implementation provisions, but it does not account for the degree to which the actors actually implement the settlement. While these research questions require different measures, our measurement strategy is appropriate for a large number of questions pertaining to an agreement's underlying strength.

Bayesian IRT measurement models have been used to study many phenomena in international relations, and the full probability model approach we employ allows these

approaches to be used even when data are scarce. Smith & Spaniel (2018) use only 16 observable indicators of nuclear capability compared to the 18 provisions we use in our analysis, but they have 8,806 country-year observations to draw from. Benson & Clinton (2016) include 489 observations in their Bayesian measurement model of alliance scope and depth. In comparison, we measure the strength of only 111 peace agreements due to limited data availability. Yet due to the additional information contained in the predictors of agreement strength included in the full probability model, we are able to obtain stable estimates of agreement strength despite the small sample size. When lots of data are available, the additional effort required of the full probability model may not be warranted, but when observations are few, the increases in measurement validity make it worthwhile.

It is also important to highlight the shortcomings of our approach. The use of a full probability model that includes predictors in addition to a measurement model allows us to produce stable estimates of agreement strength despite our small sample size. However, this means that our measurements cannot easily be included in other analyses as response or explanatory variables. Instead, researchers must estimate a full probability model using their predictors of interest. The full probability model can be computationally costly as the 90,000 iterations used to estimate Model 6 take approximately a half hour on a Linux cluster using 10 CPU cores and 220GB of RAM. These issues do not affect standard IRT models, which can be used when data are sufficiently plentiful. When data are scare, as with our sample of 111 agreements, we believe that the ability to reliably estimate latent constructs outweighs the added computational burden. Replication materials for this article include Stan code that researchers can use to estimate a full probability model for their topic of interest, lowering the barrier to using this technique.

Based on our results here, we make some basic methodological recommendations for researchers wishing to use item response theory to measure the strength of peace

agreements. First, if there are any agreements that are especially relevant to your theoretical argument, do not select them as identification restrictions. Agreements used to anchor the latent scale can shift greatly in the ranking of agreements when compared to a model where they are not selected as an identification restriction. However, agreements that are not chosen as identification restrictions rarely move more than five places in the ranking under different identification strategies. Second, use a full probability model instead of separate IRT and regression models. While this does lead to models which require more MCMC samples, the estimates produced by them incorporate more information about the phenomena at hand and lead to better predictive accuracy when used in other analyses, as seen in our study of agreement strength and duration.

The ability to reliably measure the strength of such small numbers of agreements opens up many new opportunities to ask questions we could not previously evaluate systematically. Are stronger peace agreements more likely to see full implementation (Joshi & Darby 2013, Joshi, Quinn & Regan 2015) of their various elements? Do different types of mediator leverage (Reid 2017*b*) lead to stronger or weaker agreements? Do biased mediators (Svensson 2009) lead to stronger agreements than unbiased ones? Do multilateral mediation efforts (Böhmelt 2012) produce stronger agreements than unilateral ones? While we focus on intrastate conflicts due to the wealth of mediation data in the CWM data, analyses which employ explanatory variables also available for interstate conflicts can utilize all 216 agreements in the UCDP Peace Agreements Data. Such analyses could explore whether certain factors better explain agreement strength in each type of conflict.

Further research is needed to investigate the conditions under which mediation negatively impacts agreement strength. If future work discovers that mediation can weaken agreements, this would suggest that merely solving the time inconsistency problem

(Beardsley 2008) would not lead to stronger mediated agreements. This is just one example of a research question we would be unable to answer if we could not directly measure agreement strength and instead were forced to rely on duration as a proxy. By being able to measure the strength of peace agreements irrespective of their eventual success or failure, we can increase the range of questions we can ask, leading to a better understanding of conflict termination overall. Bayesian IRT can be used to better measure existing concepts when observations and observable indicators are few (Smith & Spaniel 2018), but this paper shows that it can also be used to ask questions we otherwise would not be able to.

References

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

Anna, Cara. 2015. "U.S. Pursues New UN Sanctions on South Sudan as Deal Unsigned." *The Orange County Register* .

Arbetman, Marina & Jacek Kugler. 1997. *Political Capacity and Economic Behavior*. Boulder, Colo.: Westview Press.

Armstrong, David, Ryan Bakker, Royce Carroll, Christopher Hare, Keith T. Poole & Howard Rosenthal. 2014. *Analyzing Spatial Models of Choice and Judgment with R*. Boca Raton, FL: CRC Press.

Badran, Ramzi. 2014. "Intrastate Peace Agreements and the Durability of Peace." *Conflict Management and Peace Science* 31(2):193–217.

Bafumi, Joseph, Andrew Gelman, David K. Park & Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13(2):171–187.

Bakker, Ryan, Daniel W Hill & Will H Moore. 2016. "How Much Terror? Dissidents, Governments, Institutions, and the Cross-National Study of Terror Attacks." *Journal of Peace Research* 53(5):711–726.

Bapat, Navin A. & Bo Ram Kwon. 2015. "When Are Sanctions Effective? A Bargaining and Enforcement Framework." *International Organization* 69(01):131–162.

Bearce, David H. & Daniel C. Tirone. 2010. "Foreign Aid Effectiveness and the Strategic Goals of Donor Governments." *The Journal of Politics* 72(3):837–851.

Beardsley, Kyle. 2008. "Agreement without Peace? International Mediation and Time Inconsistency Problems." *American Journal of Political Science* 52(4):723–740.

Beardsley, Kyle C., David M. Quinn, Bidisha Biswas & Jonathan Wilkenfeld. 2006. "Mediation Style and Crisis Outcomes." *Journal of Conflict Resolution* 50(1):58–86.

Benson, Brett V. & Joshua D. Clinton. 2016. "Assessing the Variation of Formal Military Alliances." *Journal of Conflict Resolution* 60(5):866–898.

Böhmelt, Tobias. 2012. "Why Many Cooks If They Can Spoil the Broth? The Determinants of Multiparty Mediation." *Journal of Peace Research* 49(5):701–715.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. "Stan A Probabilistic Programming Language." *Journal of Statistical Software* 76(1).

Carroll, Robert & Brenton Kenkel. \noop{3001}Forthcoming. "Prediction, Proxies, and Power." *American Journal of Political Science* .

Clinton, Joshua D. & David E. Lewis. 2008. "Expert Opinion, Agency Characteristics, and Agency Preferences." *Political Analysis* 16(1):3–20.

Clinton, Joshua, Simon Jackman & Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

Copelovitch, Mark, Christopher Gandrud & Mark Hallerberg. 2018. "Financial Data Transparency, International Institutions, and Sovereign Borrowing Costs." *International Studies Quarterly* 62(1):23–41.

Cunningham, David E. 2006. "Veto Players and Civil War Duration." *American Journal of Political Science* 50(4):875–892.

DeRouen, Karl & Jacob Bercovitch. 2012. Trends in Civil War Mediation. In *Peace and Conflict 2012*, ed. J. Joseph Hewitt, Jonathan Wilkenfeld, Ted Robert Gurr & Birger Heldt. College Park, Md.]; Boulder [Colo.]; Stockholm, Sweden: Center for International Development and Conflict Management, University of Maryland ; Paradigm Publishers ; Folke Bernadotte Academy.

DeRouen, Karl, Jacob Bercovitch & Paulina Pospieszna. 2011. "Introducing the Civil Wars Mediation (CWM) Dataset." *Journal of Peace Research* 48(5):663–672.

DiStefano, Christine, Min Zhu & Diana Mindrila. 2009. "Understanding and Using Factor Scores: Considerations for the Applied Researcher." *Practical Assessment, Research & Evaluation* 14(20).

Dreher, Axel, Jan-Egbert Sturm & James Raymond Vreeland. 2009. "Development Aid and International Politics: Does Membership on the UN Security Council Influence World Bank Decisions?" *Journal of Development Economics* 88(1):1–18.

Dumo, Denis. 2015. "South Sudan President Signs Peace Deal despite Concerns." *Reuters* .

Dür, Andreas, Leonardo Baccini & Manfred Elsig. 2014. "The Design of International Trade Agreements: Introducing a New Dataset." *The Review of International Organizations* 9(3):353–375.

Fariss, Christopher J. 2014. "Respect for Human Rights Has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(2):297–318.

Fortna, Virginia Page. 2003. "Scraps of Paper? Agreements and the Durability of Peace." *International Organization* 57(2):337–372.

Gartner, Scott Sigmund. 2011. "Signs of Trouble: Regional Organization Mediation and Civil War Agreement Durability." *The Journal of Politics* 73(2):380–390.

Gartner, Scott Sigmund & Jacob Bercovitch. 2006. "Overcoming Obstacles to Peace: The Contribution of Mediation to Short-Lived Conflict Settlements." *International Studies Quarterly* 50(4):819–840.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg & Håvard Strand. 2002. "Armed Conflict 1946-2001: A New Dataset." *Journal of Peace Research* 39(5):615–637.

Gray, Julia & Jonathan B. Slapin. 2012. "How Effective Are Preferential Trade Agreements? Ask the Experts." *The Review of International Organizations* 7(3):309–333.

Greig, J. Michael. 2005. "Stepping Into the Fray: When Do Mediators Mediate?" *American Journal of Political Science* 49(2):249–266.

Harbom, Lotta, Stina Högbladh & Peter Wallensteen. 2006. "Armed Conflict and Peace Agreements." *Journal of Peace Research* 43(5):617–631.

Hare, Christopher, David A. Armstrong, Ryan Bakker, Royce Carroll & Keith T. Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Hartzell, Caroline A. 1999. "Explaining the Stability of Negotiated Settlements to Intrastate Wars." *Journal of Conflict Resolution* 43(1):3–22.

Hartzell, Caroline & Matthew Hoddie. 2003. "Institutionalizing Peace: Power Sharing and Post-Civil War Conflict Management." *American Journal of Political Science* 47(2):318–332.

Hartzell, Caroline, Matthew Hoddie & Donald Rothchild. 2001. "Stabilizing the Peace after Civil War: An Investigation of Some Key Variables." *International Organization* 55(1):183–208.

Hollyer, James R., B. Peter Rosendorff & James Raymond Vreeland. 2014. "Measuring Transparency." *Political Analysis* 22(4):413–434.

Jackman, Simon. 2000. "Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation." *Political Analysis* 8(4):307–332.

Janssen, Rianne, Francis Tuerlinckx, Michel Meulders & Paul de Boeck. 2000. "A Hierarchical IRT Model for Criterion-Referenced Measurement." *Journal of Educational and Behavioral Statistics* 25(3):285–306.

Jo, Dong-Joon & Erik Gartzke. 2007. "Determinants of Nuclear Weapons Proliferation." *Journal of Conflict Resolution* 51(1):167–194.

Joshi, Madhav, Jason Michael Quinn & Patrick M Regan. 2015. "Annualized Implementation Data on Comprehensive Intrastate Peace Accords, 1989–2012." *Journal of Peace Research* 52(4):551–562.

Joshi, Madhav & John Darby. 2013. "Introducing the Peace Accords Matrix (PAM): A Database of Comprehensive Peace Agreements and Their Implementation, 1989–2007." *Peacebuilding* 1(2):256–274.

Kadera, Kelly M. 1998. "Transmission, Barriers, and Constraints: A Dynamic Model of the Spread of War." *Journal of Conflict Resolution* 42(3):367–387.

Koford, Kenneth. 1989. "Dimensions in Congressional Voting." *American Political Science Review* 83(3):949–962.

Kugler, Jacek & Ronald L. Tammen. 2012. *The Performance of Nations*. Lanham, Md.: Rowman & Littlefield.

Little, Roderick & Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. ed. Hoboken, N.J.: Wiley.

Marshall, Monty G., Ted Robert Gurr & Keith Jaggers. 2014. POLITY IV PROJECT: Dataset Users' Manual. Codebook Center for Systemic Peace.

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9(4):538–558.

Morgan, T. Clifton, Navin Bapat & Valentin Krustev. 2009. "The Threat and Imposition of Economic Sanctions, 1971—2000*." *Conflict Management and Peace Science* 26(1):92–110.

Morgan, T. Clifton, Navin Bapat & Yoshiharu Kobayashi. 2014. "Threat and Imposition of Economic Sanctions 1945–2005: Updating the TIES Dataset." *Conflict Management and Peace Science* 31(5):541–558.

Nichols, Michelle. 2015. "U.S. Proposes United Nations Arms Embargo on South Sudan." *Reuters* .

Olson, Marie & Frederic S. Pearson. 2002. "Civil War Characteristics, Mediators, and Resolution." *Conflict Resolution Quarterly* 19(4):421–445.

Pemstein, Daniel, Stephen A. Meserve & James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426–449.

Pickering, Jeffrey & Emizet F. Kisangani. 2009. "The International Military Intervention Dataset: An Updated Resource for Conflict Scholars." *Journal of Peace Research* 46(4):589–599.

Poole, Keith & Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Poole, Keith T. & Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357–384.

Poole, Keith T., Howard Rosenthal & Kenneth Koford. 1991. "On Dimensionalizing Roll Call Votes in the U.S. Congress." *American Political Science Review* 85(3):955–976.

R Core Team. 2018. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing.

Rasch, G. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press. OCLC: 6278633.

Regan, Patrick M. 2002. "Third-Party Interventions and the Duration of Intrastate Conflicts." *Journal of Conflict Resolution* 46(1):55–73.

Reid, Lindsay. 2017*a*. "Catering to the Conflict: An Assessment of the Durability of Context-Specific Peace Agreements.".

Reid, Lindsay. 2017*b*. "Finding a Peace That Lasts: Mediator Leverage and the Durable Resolution of Civil Wars." *Journal of Conflict Resolution* 61(7):1401–1431.

Schnakenberg, Keith E. & Christopher J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.

Sheng, Yanyan & Christopher K. Wikle. 2008. "Bayesian Multidimensional IRT Models With a Hierarchical Structure." *Educational and Psychological Measurement* 68(3):413–430.

Singer, J. David. 1988. "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816–1985." *International Interactions* 14(2):115–132.

Singer, J. David, Stuart A. Bremer & John Stuckey. 1972. Capability Distribution, Uncertainty, and Major Power War, 1820–1965. In *Peace, War, and Numbers*, ed. Bruce Russett. Beverly Hills: Sage Publications pp. 19–48.

Smith, Bradley C. & William Spaniel. 2018. "Introducing *v*-CLEAR: A Latent Variable Approach to Measuring Nuclear Proficiency." *Conflict Management and Peace Science* p. 0738894217741619.

Stan Development Team. 2017. "Rstan: R Interface to Stan.".

Svensson, Isak. 2009. "Who Brings Which Peace? Neutral versus Biased Mediation and Institutional Peace Arrangements in Civil Wars." *Journal of Conflict Resolution* 53(3):446–469.

Themnér, Lotta & Peter Wallensteen. 2014. "Armed Conflicts, 1946–2013." *Journal of Peace Research* 51(4):541–554.

Touval, Saadia & I. William Zartman. 1985. Introduction: Mediation in Theory. In *International Mediation in Theory and Practice*, ed. Saadia Touval & I. William Zartman. Boulder, CO: Westview Press.

Treier, Shawn & Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–217.

van Buuren, Stef & Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45(3).

Wehr, Paul & John Paul Lederach. 1991. "Mediating Conflict in Central America." *Journal of Peace Research* 28(1):85–98.

Werner, Suzanne & Amy Yuen. 2005. "Making and Keeping Peace." *International Organization* 59(2):261–292.

World Bank. 2018. World Development Indicators 2018. Technical report Washington, D.C.: .