Measuring Peace Agreement Strength in Civil War*

Rob Williams†

August 20, 2021

Abstract

Scholars have long been interested in the strength of peace agreements, but conventional measurement practices are often subjective and endogenous to other post-conflict processes. Even more recent efforts rely on conflict-specific information that limits the usability of the resulting estimates. I introduce an approach for measuring strength ex ante by relying on the content of agreements alone. I explore multiple variations of this approach and empirically evaluate the estimates produced by each. Drawing on three different datasets of peace agreements signed in all UCDP/PRIO civil conflicts from 1975-2018, the estimates capture the strength of peace agreements at the time of signing. The estimates show descriptively that stronger peace agreements tend to be signed in more intractable conflicts, suggesting that a selection effect may be at play in the process of agreement signing and duration.

Word count: 11,037

The strength of peace agreements is an elusive concept that lies under the surface of much scholarship on conflict management and resolution. Numerous studies explore the relationship between specific provisions of peace agreements in civil conflict and the likelihood of renewed violence (Hartzell, Hoddie, and Rothchild 2001; Hartzell and Hoddie 2003; Werner and Yuen 2005; Matanock 2017; Reid 2017). Left unsaid in this research is that idea that if these provisions can make an agreement last longer, they also make that agreement stronger. Building on a conceptualization of peace agreement strength introduced by Fortna (2003), this manuscript explores multiple different ways of measuring agreement strength. Using the Bayesian latent variable model in Williams et al. (2021) as a starting point, this manuscript develops an approach for measuring agreement strength and empirically evaluates the accuracy of multiple potential variations within this scheme. This framework allows researchers to generate measures of agreement strength that best fit their specific questions.

*I thank Mike Kenwick, Jeff Carter, Marcella Morris, Guillermo Rosas, and Layna Mosley for helpful comments.
†Postdoctoral Research Fellow, Washington University in St. Louis, rob.williams@wustl.edu, jayrobwilliams.com

This approach offers multiple advantages over existing measures of peace agreement strength. Picking one or two theoretically motivated provisions ignores the influence of all other provisions and ignores the fact that some provisions may be more relevant to some conflicts than others. Creating an additive index of agreement provisions as Werner and Yuen (2005) do assumes that all provisions contribute equally to agreement strength. Joshi and Quinn (2015) instead use an additive index of agreement provisions to capture the degree of reform involved in an agreement, but this makes the same assumption that all provisions entail an equal amount of reform to a post-conflict society.[1] Finally, Williams et al. (2021) employ a similar latent variable model but rely on conflict-level information that prevents their measure from being employed in any analysis involving the same conflict-level variables. As this conflict-level information includes factors such as the underlying incompatibility in a conflict, this severely limits the applicability of their measure in other contexts. This approach builds on this earlier work to deliver a nuanced and broadly-applicable measure of peace agreement strength that captures the strength of peace agreements at the time of signing.

The latent variable model provides a single measure of peace agreement strength based on the characteristics of each agreement and allows direct comparison between different agreements. In addition, the model also offers insights into the relationship between individual provisions and the strength of agreements. In the same way that scholars of conflict management have begun to disaggregate power sharing to explore the effect of different provisions on agreement durability (Mattes and Savun 2009, 741), this approach facilitates investigating the individual impact of different provisions on agreement strength while accounting for their simultaneous presence or absence with other provisions. The implementation in this manuscript draws on data from the UCDP Peace Agreement Dataset (Harbom, Högbladh, and Wallensteen 2006), PA-X (Bell and Badanjak 2019), and the Peace Accords Matrix (Joshi, Quinn, and Regan 2015) to the measure.

This procedure provides a comprehensive measure of peace agreement strength at the time of signing. It does so by relying only on information contained within the document itself. Statistically this means that the scores can be used in any regression analysis without concerns of endeogeneity because they do not incorporate any information about the conflict external to the agreements themselves. This agreement-focused approach means that the framework can be employed in future research to try and isolate the independent

[1]The presence or absence of agreement provisions is encoded in binary form in peace agreement data, so a sum of provisions entails treating each provision as contributing equally to the final score because each provision has a numerical value of one.

effect of peace agreements as institutions, separating out the role of conflict-level factors. I demonstrate the utility of this approach by generating a measure of agreement strength for agreements signed in civil conflicts from 1975-2018 and using it to show that agreement strength may be subject to a selection effect where more intractible conflicts are more likely to end with strong agreements.

1 Peace agreement strength

There is a long-running debate in international relations over whether institutions have an independent effect on behavior or are endogenous reflections of the underlying preferences of involved actors (Keohane 1988).[2] In the study of conflict management, this debates centers on whether peace agreements have an independent effect on the likelihood of conflict recurrence, or whether they are ephemeral "scraps of paper" whose apparent influence is driven by underlying forces (Fortna 2003). To facilitate further investigation of this debate, the framework that I develop does not require the use of any external information to estimate agreement strength.

Given the well-documented differences between conflicts fought over government and territory (Buhaug 2006; Cederman, Buhaug, and Rød 2009), including information on which type of conflict an agreement was signed in would likely generate a measure of agreement strength that has higher predictive accuracy than one that only includes information from the agreement itself. Such a strategy would account for unmeasured covariation between agreements in each type of conflict due to omitted variables. However, this measure would include both endogenous and exogenous aspects of peace agreements and would be unable to evaluate whether peace agreements have independent effect on post-conflict outcomes.

For this reason, I develop a baseline framework that only includes information on the specific provisions within peace agreements as signed documents. As another example, agreements signed as part of mediation processes may be weaker due to a narrow-focused desire to achieve an agreement sooner at the expense of a stronger agreement (Svensson 2009; Williams et al. 2021). However, including the presence or absence of mediation in the measure would mean that the estimates would be unsuitable for any regression analysis that used mediation as a predictor. By only drawing on information contained within the text of peace

[2]For work in the area of IMF treaty compliance see Simmons (2000), von Stein (2005), and Simmons and Hopkins (2005); for work in the area of human rights treaty compliance see Hathaway (2002) and Simmons (2009).

agreements, this framework can be used in any analysis involving peace agreements, whereas a measure that incorporated more general information about the conflict could not.

While the framework does not require the inclusion of conflict-level information, it does not necessarily preclude it either. If the research question at hand does not require controlling for certain characteristics of the conflict, then they could be included in the measure to potentially generate more accurate estimates of agreement strength. I explore this possibility when empirically evaluating different implementations of the model to explore the utility of including conflict-level information.

The discussion of the independent effect of peace agreements on phenomena such as conflict recurrence or public health outcomes in the aftermath of conflict implies that there is a causal effect of peace agreements on these outcomes. We wish to know how the outcome would change not only in the absence of an agreement, but also if the substance of that agreement were different while all other relevant factors remained constant. The concept of peace agreement strength in the conflict management literature is synonymous with this independent effect of agreements; an agreement that has a larger effect on the durability of peace is substantively stronger than one with a smaller impact. Fortna operationalizes the strength of a peace agreement as the "number and extent of the measures implemented as part of a cease-fire" (2003), suggesting that individual measures contribute to the effect of an agreement. Within this broad understanding of agreement strength, we can identify two clear categories of measures.

The first — "conflict resolution provisions" — are "stipulations in peace agreements aiming to resolve the basic incompatibilities" (Svensson 2007, 241). The provisions are intended to identify compromises that can prevent the return to hostilities in the future. Examples include power sharing arrangements to guarantee representation in legislative bodies to former rebel movements (Hartzell and Hoddie 2003) or the granting of regional autonomy in the wake of conflict (Cederman et al. 2015). If conflict resolution provisions aim to resolve fundamental incompatibilities underlying conflict, then agreements that contain more provisions should address more incompatibilities and reduce the incentive for renewed conflict in the future. In addition to addressing multidimensional grievances, multiple conflict resolution provisions can provide a form of redundancy whereby even if one fails, parties can turn to others to keep their rivals in check (Hartzell and Hoddie 2019, 643).

Conflict prevention provisions are smaller in scope and more narrowly focused on preventing the resumption of hostilities due to miscommunication or accident. International peacekeeping missions decrease the likelihood of renewed conflict after an agreement by reducing uncertainty through monitoring, raise the cost of returning to conflict, and can manage accidents to prevent them from escalating into renewed hostilities (Fortna 2008; Ruggeri, Dorussen, and Gizelis 2017/ed). Other provisions, like transitional justice institutions, are explicitly designed to address the consequences of conflict and not the cause of the conflict. While power sharing arrangements function by addressing root causes of conflict, they are more effective when they actually see implementation (Jarstad and Nilsson 2008; Ottmann and Vüllers 2015). Efforts to ensure the full implementation of the stipulations of an agreement, such as detailed timelines or provisions for external review of post-conflict measures, are thus also conflict prevention provisions (Joshi and Quinn 2017). While these provisions do not address the root of the conflict, they have a similarly pacifying effect on the likelihood of future conflict.

While each type of provision functions through different causal pathways, they both work to decrease the likelihood of renewed conflict. The method relies on both types of provisions to generate a comprehensive measure of peace agreement strength. Both conflict resolution and prevention provisions are specific pieces of language within agreements that stipulate concrete actions that parties will take after the cessation of hostilities (Harbom, Högbladh, and Wallensteen 2006).

2 Measurement strategy

The basis of this approach to measuring peace agreement strength is to treat it as a latent variable measured using the two parameter item response model (Rasch 1980). This approach models the provisions included in peace agreements as probabilistic functions of the latent strength of the agreement, with stronger agreements being more likely to exhibit provisions. The functional form of this relationship is assumed to be the logistic function, and the model is given by

$$\Pr(y_{ij} = 1) = \text{logit}^{-1}[\gamma_j(\theta_i - \alpha_j)] \tag{1}$$

where i indexes agreements and j indexes provisions. θ_i is the latent strength of an agreement, and higher θ_i values are associated with a higher likelihood of observing a given $y_{ij} = 1$. However, this likelihood is not equal for all provisions and the remaining two parameters control the location and shape of the logistic curve for each provision. The difficulty parameter α_j serves as a baseline and reflects the probability of observing $y_{ij} = 1$ when θ_i is equal to the mean of the latent strength $\bar{\theta}$, which is also the value of θ_i at which there is an equal probability of observing $y_{ij} = 0$ or $y_{ij} = 1$. Lower values mean that an agreement need not be particularly strong for us to observe $y_{ij} = 1$. The discrimination parameter γ_j controls the slope of the logistic curve, which corresponds to how well a given provision discriminates between weak and strong agreements. When γ_j is low, the slope of the curve is low, and a shift in θ_i results in only a minimal change to $\Pr(y_{ij} = 1)$, so there is a large region of uncertainty about the strength of an agreement. In contrast, when α_j is high, the region of uncertainty is small, and only minimal changes in θ_i are needed to shift $\Pr(y_{ij} = 1)$ from ≈ 0 to ≈ 1.

A high α value and a low γ would indicate a provision that is associated with strong agreements, but does a poor job separating stronger and weaker agreements. This means that overall agreements with this provision will be stronger, but that large increases in agreement strength only marginally increase the probability of observing that agreement. As observers not privy to the data generating process behind real world data, a low γ_j estimate tells us that given two agreements with otherwise equal provisions, the agreement with $y_{ij} = 1$ may not actually be stronger than the agreement with $y_{ij} = 0$. Below, I consider five alternative implementations of this model as potential candidates for the most broadly applicable measure of agreement strength.

2.1 Baseline model

The model with priors and hyperpriors is presented below:

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, 1) \tag{2}$$
$$\boldsymbol{\gamma} \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma) \tag{3}$$
$$\boldsymbol{\alpha} \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha) \tag{4}$$
$$\boldsymbol{\beta} \sim \mathcal{N}(0, 1) \tag{5}$$
$$\mu_\gamma, \mu_\alpha \sim \mathcal{N}(0, 1) \tag{6}$$
$$\sigma_\alpha, \sigma_\gamma \sim \exp(1) \tag{7}$$

The prior mean of θ is defined as a linear combination of an agreement's *type*, and a set of coefficients β. *Agreement type* denotes whether an agreement is a process, partial, or full agreement. Full agreements reflect attempts to settle the entire incompatibility in a conflict, partial ones address part of the incompatibility, while process agreements indicate merely "initiate a process to settle the incompatibility" (Harbom, Högbladh, and Wallensteen 2006, 622). This variable is coded -1 for process agreements, 0 for partial agreements, and 1 for full agreements.

This model is similar to the one in Williams et al. (2021), but with two key improvements. Their identification strategy uses a weak and a strong agreement to orient and scale the latent measurement of peace agreement strength and the rank of these anchor agreements in the range of latent strengths is highly unstable and subject to large shifts when different agreements are used as anchor points. Their model necessitates the inclusion of information about a number of conflict-level factors in civil war, such as economic sanctions or mediation attempts, meaning that their measure cannot be used to explore the independent effect of peace agreement strength. They also include a conflict-level random intercept which indirectly includes conflict-level information by sharing information between agreements signed in the same conflicts. Neither of these limitations apply to this model, as it does not include any conflict-level information and employs a much more general identification strategy detailed below.

Williams et al. (2021) find that their measure is positively correlated with *agreement type*, and I leverage this relationship to identify the model. Constraining the coefficient on *agreement type* to be positive places weak and strong agreements on opposite sides of the likelihood surface, solving the reflection invariance problem and and ensuring that strong agreements will have positive estimates and weak agreements negative ones (Bafumi et al. 2005, 176–79).

$$\beta_{type} > 0 \tag{8}$$

However, *agreement type* does not cleanly separate strong and weak agreements.[3] The model thus includes a second identification restriction where the discrimination parameters are constrained to be positive under

[3]Using only *agreement type* as an identification restriction results in an unidentified model; see the Supplemental Information for details.

the assumption that the presence of a provision always signals a stronger agreement.[4] Note that while a discrimination parameters are constrained to be strictly positive, they can be close to zero, meaning that their inclusion does not meaningfully contribute to the strength of an agreement. Even if the targeted incompatibility for a provision is not present in a conflict, the inclusion of that provision would thus fail to increase the strength of an agreement, rather than decrease it.

$$\gamma > 0 \tag{9}$$

2.2 Full model

The baseline model assumes that data on peace agreements only come from one source, as it cannot handle any missing data in the provisions. However, there are multiple different sources of data on the provisions within peace agreements (Harbom, Högbladh, and Wallensteen 2006; Joshi, Quinn, and Regan 2015; Bell and Badanjak 2019), with imperfectly overlapping coverage between them. The baseline model would not allow the inclusion of information from any source that does not provide data on provisions for every single agreement in the analysis. This shortcoming is problematic because different datasets focus on different aspects of peace agreements, so omitting a dataset may mean throwing out useful information. To overcome this limitation, I follow Carter and Smith Jr. (2020) and extend the baseline model to allow provisions with imperfect data coverage to inform the prior on θ alongside *agreement type* as part of $\mathbf{X}$ in Equation 2.

2.3 Conflict-level model

One natural extension is to include additional conflict-level covariates, such as the *incompatibility* under dispute or whether *ethnic* cleavages motivate the violence, as part of $\mathbf{X}$ in Equation 2. This decision narrows the applicability of the resulting measure because it would not be suitable to use in an analysis including either of these variables on the other side of the regression equation. In cases where the conflict-level information is not relevant to the question under study, this may not pose an issue. I estimate a version of the model that includes XXXXX as covariates in $\mathbf{X}$ to see if they increase the predictive accuracy of the model

[4] While this is sufficient to identify the model, convergence requires over 100,000 MCMC iterations. To aid with convergence μ_γ is also constrained to be > 0.

when they are not also included as control variables.

2.4 Robust model

One potential concern with the model in Equation 1 is that it assumes that the relationship between a given provision and agreement strength is globally homogeneous and does not depend on local circumstances. To give a concrete example, the model assumes that the presence of disarmament provisions communicates the same information about the strength of an agreement regardless of whether a conflict is fought over territorial or governmental incompatibilities. To test whether this assumption is problematic, I consider an extension of the two parameter item response model that introduces a third error parameter, $\boldsymbol{\epsilon}$,

$$\Pr(y_{ij} = 1) = \epsilon_0 + (1 - \epsilon_0 - \epsilon_1)\text{logit}^{-1}[\gamma_j(\theta_i - \alpha_j)] \tag{10}$$

which accounts for the possibility that any given provision is not actually related to the issue under dispute, and was thus included by 'accident.' In the educational context, $\boldsymbol{\epsilon}$ is traditionally used to model guessing on exams (Johnson and Albert 1999, 204–5; Bafumi et al. 2005, 178–79). In the realm of peace agreements, ϵ_0 represents the probability that a weak agreement may include a provision irrelevant to the conflict at hand and misleadingly appear stronger than it is, while ϵ_1 is the probability that a strong agreement fails to include a relevant provision. The effect of $\boldsymbol{\epsilon}$ is to set a 'floor' and 'ceiling' on the logistic curve.

The prior on $\boldsymbol{\epsilon}$ is chosen to be $\mathcal{U}(0, 0.1)$ as any value above 0.1 would raise concern about the appropriateness of fitting an IRT model to the data (Bafumi et al. 2005, 179). The error parameter ϵ_0 is estimated to be 0.01 and ϵ_1 is estimated to be 0.07, so the minimum probability of observing a given provision is 0.01 and the maximum is 0.93. Given these relatively low error rates, the three parameter IRT model in Equation 10 has a classification accuracy (84.35%) indistinguishable from the two parameter model (84.27%). Given the equal predictive accuracy, I use the more parsimonious two parameter model.

2.5 Differential item functioning model

One way to explicitly model the fact that different provisions are more or less relevant in different conflicts due to varying issue saliences across conflicts would be to allow for differential item functioning in the model. This would let $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ vary by conflict to capture the fact that specific provisions are more important to resolving different disputes. In conflicts where cultural issues are prominent, cultural freedoms and local governance provisions will require a stronger agreement to observe and will better differentiate between strong and weak agreements than in conflicts that are more governmental in nature because they directly address the underlying disagreement. However, this would introduce 3,528 new parameters with no corresponding increase in data, requiring stronger identification restrictions that would narrow the applicability of the scores.

A more feasible approach is to evaluate whether we observe differential item functioning by incompatibility. Territorial conflicts can be more difficult to resolve than governmental ones (M. Toft 2003) due to the strategic or identity value of territory (M. D. Toft 2014), so we might expect difficulty parameters to be higher in these conflicts. This model includes two sets of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ vectors, one for conflicts over territory and one for conflicts over government, which accounts for the fact that provisions such as federalism and local power sharing may contribute more to resolving territorial incompatibilities than governmental ones. Similarly, rebels engaged in territorial conflicts may not be placated by offers of integration into the civil service. The model in Equation 1 becomes

$$\Pr(y_{ijk} = 1) = \text{logit}^{-1}[\gamma_{jk}(\theta_{ik} - \alpha_{jk})] \tag{11}$$

where k indexes the incompatibility in conflict i (governmental or territorial). The priors on the difficulty and discrimination parameters are similarly indexed.

$$\boldsymbol{\gamma}_k \sim \mathcal{N}(\mu_{\gamma_k}, \sigma_{\gamma_k}) \tag{12}$$
$$\boldsymbol{\alpha}_k \sim \mathcal{N}(\mu_{\alpha_k}, \sigma_{\alpha_k}) \tag{13}$$

This yields four different forms of the model: a baseline model, the full model with included information from other peace agreement datasets, a model that also includes conflict-level information, a robust model

that allows for the inclusion of irrelevant provisions, and a model with differential item functioning to account for the incompatibility underlying the conflict. I next assess the external predictive validity of each specification, before moving onto to the internal validity of the model with differential item functioning, which has the best predictive accuracy.

3 The data

My evaluation of the framework employs the UCDP Peace Agreement Dataset version 19.1 (Pettersson, Högbladh, and Öberg 2019), which contains information on the provisions contained in 324 unique peace agreements from 1975-2018.[5] The data encompass both conflict resolution and c onflict prevention provisions. Examples of conflict resolution provisions include power sharing and territorial autonomy arrangements, while disarmament and withdrawal of foreign forces number among the conflict prevention provisions. Table 1 presents all 28 provisions in the data.

Ceasefire	Elections	Referendum	Prisoner Release
Military Integration	Interim Government	Local power Sharing	National Reconciliation
Disarmament	National Talks	Regional Development	Right of Return
Withdrawal	Power Sharing	Cultural Freedoms	Reaffirmation
Political Parties	Territorial Autonomy	Border Demarcation	Peacekeeping
Government Integration	Federalism	Local Governance	Gender Provisions
Civil service Integration	Independence	Amnesty for Rebels	Implementation Commission

Table 1: Peace agreement provisions in the UCDP peace agreements data

All of these provisions are plausibly relevant to the strength of peace agreements, but it is unlikely that they all contribute equally to the strength of an agreement. The measurement approach detailed in Section 2 incorporates all provisions, but uses patterns of co-occurrence to learn how individual provisions contribute differentially to agreement strength. This technique strikes a balance between relying on previously-published studies to identify a subset of agreements to select as indicators of strength, and creating an additive index of all provisions to capture strength.[6]

One provision in particular warrants in-depth discussion. The implementation commission provision denotes whether an agreement stipulated a committee or commission to oversee the implementation of the agreement (Harbom, Högbladh, and Wallensteen 2006). This variable is a provision included in the document,

[5]Three agreements are signed to terminate multiple conflicts. The SI details how these agreements are handled.
[6]The outlining provision is omitted following exploratory analyses. See the Supplemental Information for full details.

and is not a measure of any post-signing implementation activities or lack thereof. Accordingly, including it does not risk introducing post-treatment bias due to the endogeneity of any post-signing activities to the agreement itself and the negotiation process that produced the agreement. The degree of implementation realized in a post-conflict society has a large impact on the eventual duration of the peace (Joshi and Quinn 2017), so it is important to measure whether an agreement includes arrangements designed to facilitate full implementation in the future.

Inspecting the provisions in specific agreements can give some insight into the strength of the agreements. Figure 1 presents provisions in four different agreements:

- The Lancaster House Agreement that ended the Rhodesian Bush War in 1979
- The comprehensive peace agreement between the government of Colombia and the FARC in 2016
- The Good Friday Agreement that ended the Troubles in 1998
- The Arusha Accords in the Rwandan Civil War in 1993

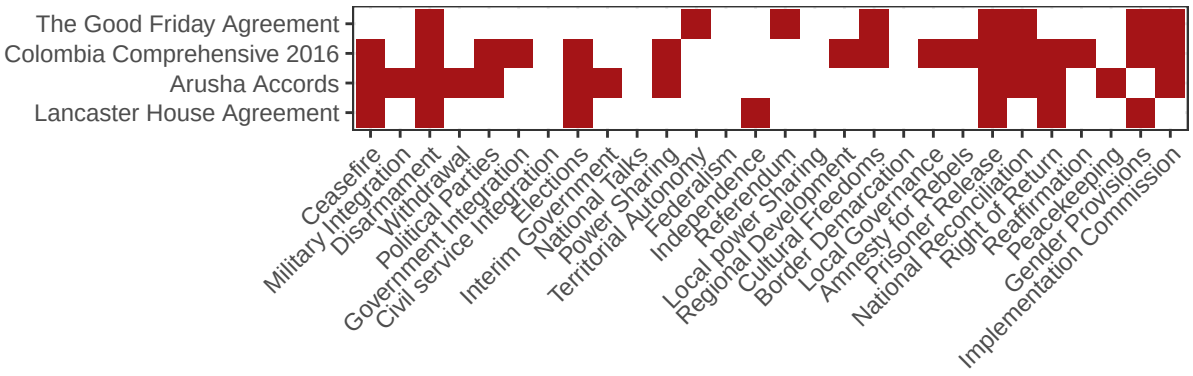

Figure 1: Provisions in various peace agreements.

The Lancaster House Agreement is one of only two agreements in the data with independence provisions, and the second longest surviving agreement in the data. The comprehensive Colombia 2016 agreement contains many post-conflict reconciliation provisions such as amnesty for former rebels, the release of prisoners, national reconciliation, and establishing a right of return for refugees and internally displaced persons. The Good Friday Agreement has rather fewer provisions and ensures disarmament of former combatants, cultural freedoms, and the holding of a national referendum on the agreement. The Arusha Accords provided for the integration of former combatants into the civil service and military, power sharing

agreements, and scheduled elections.

Each of these peace agreements is qualitatively different from the others. While all four mandate the disarmament of combatants and the release of prisoners, the differences between them are striking. The Arusha Accords focus strongly on political concessions to the Rwandan Patriotic Front while the comprehensive agreement in Colombia is oriented towards addressing the human costs of such a long-running conflict. These differences illustrate the fact that different provisions take different steps toward resolving incompatibilities or implement different mechanisms to address post-conflict commitment problems.

3.1 Additional data sources

While the UCDP Peace Agreement Dataset includes a wealth of information on the content of peace agreements, it is not exhaustive and there are multiple conspicuous oversights. Many factors that have been found to have important effects on agreement survival, such as economic power sharing and detailed implementation frameworks, are absent from the data. To address these shortcomings, the full model incorporates information from two other sources of information on peace agreements.

PA-X includes much more detailed data at more disaggregated levels than the UCDP Peace Agreements Data and records 225 different quantitative dimensions of peace agreements (Bell and Badanjak 2019). PA-X includes two types of codings for agreement provisions. Binary variables note whether a given provision was included in the agreement or not, while categorical ones code a 0 for no mentions of a provision, 1 for rhetorical mentions, 2 for substantive provisions, and 3 for detailed provisions. Examples of provisions found in PA-X but not the UCDP Peace Agreement Data include the establishment of permanent electoral commissions and judicial accountability measures.

The Peace Accords Matrix Implementation Dataset (PAM) tracks the implementation of 51 different provisions in comprehensive peace agreements over time (Joshi, Quinn, and Regan 2015). While the primary purpose of the dataset is to study factors that affect the eventual implementation of agreement provisions, the data are necessarily also a source of information on provisions included in the agreement at the time of signing. Provisions that PAM tracks that are found in neither the the UCDP Peace Agreement Data nor PA-X include detailed implementation timelines and protocols for external review of agreements.[7]

[7]The Power-Sharing Event Dataset (Ottmann and Vüllers 2015) tracks power sharing over time after an agreement is signed, but does

PA-X and PAM include both conflict resolution and conflict prevention provisions that have been shown to reduce the likelihood of future conflict. Unsurprisingly, many of the provisions mirror those in the UCDP Peace Agreements Data. Examples of provisions found in PA-X that duplicate those contained in the UCDP Peace Agreements Data include refugees and internally displaced persons, which mimics the UCDP right of return provision, and border demarcation, which is found in all three datasets.

These two datasets also include provisions that are not found in the UCDP Peace Agreements Data. Constraining power sharing institutions like independent judiciaries and guarantees of the protection civil liberties reduce the likelihood of both new and renewed conflict (Gates et al. 2016). PA-X includes multiple provisions that fall under this umbrella including judicial accountability and measures to protect specific groups. Disagreements over the distribution of natural resources can lead to conflict recurrence (Rustad and Binningsbø 2012), and PA-X provides data on both economic power sharing more broadly and natural resource arrangements specifically. Land reform can help prevent future conflict by addressing underlying inequalities and as part of a disarmament, demobilization and reintegration for former combatants on both sides (Binningsbø and Rustad 2012), and is included in PA-X. Monitoring and verification are key to reducing uncertainty about military capabilities on all sides, decreasing the likelihood of renewed conflict (Mattes and Savun 2010), and PAM notes whether an agreement contains provisions for verification mechanisms. The degree to which specified provisions are ultimately implemented has a large effect on the survival of an agreement (Joshi and Quinn 2017). PAM tracks the presence of a detailed implementation timeline and external donor support which both increase the likelihood of full implementation. See Section XXX of the SI for a full list of provisions from PA-X and PAM included in the full model.

Ideally these data from PA-X and PAM would be included in the provisions Y. Unfortunately, PA-X contains information on only 84.45% of the agreements in the UCDP Peace Agreements Data, and PAM only 9.15%. The missingness is higher than the conventionally accepted threshold of 15% for PA-X, and almost total for PAM. PA-X begins in 1990 and PAM in 1989, while PAM only covers comprehensive peace agreements, meaning the data are missing not at random, so imputing missing data before including the PA-X and PAM provisions in Equation 1 is not appropriate (Little and Rubin 2002).

not contain any provisions not already included in the existing datasets, so it is not used.

Provisions from PA-X and PAM thus cannot be included in the model alongside the provisions presented in Table 1. Instead they are included in $\mathbf{X}$ in the prior on θ as outlined in Section 2.2. Although they still add to the measure of an agreement's strength at the time of signing, they contribute less than the provisions from the UCDP Peace Agreement Data. While less than ideal, this compromise allows the model to include theoretically-relevant information it would not otherwise be able to.

3.2 Why not include interstate wars?

While negotiated settlements occur in both interstate and civil wars, they have historically been rarer in civil wars (Pillar 1983) due to heightened commitment problems relative to interstate wars (Walter 1997), although their prevalence has waxed and waned over time in response to shifts in international norms around conflict resolution (Howard and Stark 2018). Negotiated settlements may similarly be less durable in civil wars (Walter 2002) due to the need to integrate former combatants into society (Hartzell 1999; Hartzell, Hoddie, and Rothchild 2001; Hartzell and Hoddie 2003). While conflict resolution provisions like power sharing and conflict prevention provisions such as enforcement mechanisms are strong predictors of agreement success or failure in civil war, they are not in interstate conflict (Werner 1999). Some mechanisms that affect the durability of peace are only applicable in one type of conflict: foreign imposed regime change greatly increases the duration of peace after interstate conflict (Lo, Hashimoto, and Reiter 2008), but while externally imposed regime change is a plausible outcome in almost all interstate conflicts, it is not in intrastate conflict. For these reasons, it is not appropriate to pool peace agreements across types of conflict.

The UCDP Peace Agreement Dataset contains only 31 peace agreements signed in international conflicts due to the rarity of interstate conflict during the sample period. Although the model builds upon the approach introduced by Fortna (2003) to measure the strength of peace agreements in interstate conflict, this is too few observations to generate estimates with meaningful variation in a model estimated on just interstate conflict. Due to the inability to pool interstate and intrastate conflicts or generate separate estimates for interstate conflict, I only estimate the strength of agreements in civil conflict.

4 Model assessment

To compare the different ways to measure agreement strength represented by the various models in Section 2, we must look at multiple dimensions of construct validity for each approach. Before doing this, a necessary first step is to check the suitability of the identification restriction imposed on the model. Specifically, the constraint that $\beta_{type} > 0$ in Equation 8 needs to be assessed.[8] If *agreement type* is a good separator of strong and weak agreements, then the estimate for $\beta_{type} \gg 0$.

Across all five model specifications, the lowest posterior mean for β_{type} is 0.15 with a 95% credible interval of $[0.68, 1.34]$, so *agreement type* clearly serves its purpose as an identification restriction for all variations of the model.[9] With the suitability of this identification restriction established, below I compare the external and internal validity of the models.

4.1 Predictive utility

Measuring the predictive accuracy of the various models helps to assess the concurrent validity of the estimates. One way to do so is to generate predictions $\widehat{y_{ij}}$ for each provision and compare whether they match the observed values y_{ij}. I generate these predictions by evaluating Equation 1 (or 10 or 11 as appropriate) using $\widehat{\theta_i}$, $\widehat{\alpha_j}$, $\widehat{\gamma_j}$, and setting $\widehat{y_{ij}} = 1$ if the result is greater than 0.5. To measure the accuracy of each specification, I then check whether $y_{ij} = \widehat{y_{ij}}$ and take the average across all agreements i and provisions j.

	Baseline	Full	Conflict	Robust	Differential
Percent correctly classified	84.28	84.27	84.29	84.35	81.82

Table 2: Accuracy of predictions for observed indicators for different measurement model specifications

Table 2 presents these measure of accuracy. The full model correctly classifies the highest percentage of observed indicators $\mathbf{Y}$, while the model with differential item functioning correctly classifies the lowest. This decrease in accuracy suggests that the relationship between observed provisions and agreement strength may not significantly vary by incompatibility, and the extra 56 parameters that this model has to estimate reduce its accuracy. However, this quantity measures how well each model does at predicting the provisions

[8] All results are from models with 4 chains run for 20,000 iterations with 15,000 warmup iterations. Inference is performed on the 5,000 post-warmup iterations pooled across chains.

[9] Starting values for θ in the MCMC sampler are set to -3 for partial agreements, 0 for process agreements, and 3 for full agreements to speed up convergence of the chains. All other parameters are randomly initialized.

used to estimate agreement strength. To assess the usefulness of the estimates, we need to evaluate how well they do at predicting outcome of interest to scholars of peace and conflict. This approach also lets us compare their utility to that of simpler measures.

I evaluate predictive accuracy for quantities of interest: whether an agreement ends in failure and how long it survives until such (potential) failure. Agreement failure is measured as whether fighting occurs in the year(s) after an agreement is signed according to the UCDP Armed Conflict Data inclusion criteria , while survival time is measured as the difference in years between an agreement's signing and failure.[10] The former outcome is modeled with a logistic regression while the latter employs a Cox proportional hazards model. I use area under the receiver operating characteristic curve (AUC) to measure the accuracy of predicting agreement failure (Bradley 1997), and the Integrated Brier Score (IBS) for the predicted duration (Graf et al. 1999). IBS is calculated for the accuracy of predictions up to the maximum duration of agreements in the data. Both of these metrics evaluate the predictive accuracy of a model, with a higher AUC and a lower IBS indicating a more accurate model.

To gain a more complete sense of the predictive power of the models, I also carry out this process with two alternative measures of agreement strength. I test the predictive power of *agreement type*, as comprehensive agreements resolve more incompatibilities than partial agreements, and both resolve more than process ones. However, *agreement type* focuses on resolving underlying incompatibilities and does not capture steps an agreement takes to address post-conflict insecurities that may lead to renewed fighting due to commitment or enforcement problems. Creating an *additive index* of agreement provisions captures both conflict resolution and conflict prevention provisions, so I test the predictive power of one as well.[11]

Using in-sample measures of predictive power risks overfitting the model to the data, and generating poor predictions when exposed to new data (Hastie, Friedman, and Tibshirani 2009, 219–57; Ward, Greenhill, and Bakke 2010). Given the small number of peace agreements in the data, this danger is especially concerning. To alleviate these concerns, I perform 3-fold cross-validation where the data are partitioned into 3 subsets. The model is then re-fit to 2 subsets of the data and used to generate predictions for the observations in the remaining 1/3 of the sample (Efron 1983). The accuracy of these predictions is then measured using either

[10] See the SI for a full discussion of these coding procedures.

[11] Only provisions from the UCDP Peace Agreement Data are included in this additive index due to missing data issues with PA-X and PAM as discussed in Section 3.1.

AUC or IBS as appropriate, and the process is repeated 2 more times to hold out a different subset of the data each time. The metrics are then averaged across all 3 folds, yielding an overall measure of accuracy. These results are included in Tables 3 and 4 below the measures of in-sample model fit.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Agreement Type	0.10*						
	(0.04)						
Additive Index		−0.02*					
		(0.01)					
Baseline model			−0.08*				
			(0.03)				
Full model				−0.09*			
				(0.03)			
Conflict-level model					−0.09*		
					(0.03)		
Robust model						−0.09*	
						(0.03)	
Differential model							−0.10*
							(0.03)
(Intercept)	0.25*	0.55*	0.46*	0.46*	0.46*	0.46*	0.46*
	(0.09)	(0.05)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
In-sample AUC	0.43	0.41	0.41	0.40	0.40	0.40	0.39
3-fold AUC	0.35	0.40	0.67	0.67	0.67	0.67	0.71
AIC	473.74	475.53	472.79	471.28	471.75	471.40	468.93
Log Likelihood	−233.87	−234.77	−233.39	−232.64	−232.87	−232.70	−231.47
Num. obs.	328	328	328	328	328	328	328

*p < 0.05

Table 3: Logistic regression models of agreement strength and conflict recurrence

For agreement failure, agreement type performs best with an AUC of 0.43, while an additive index does best for agreement duration with an IBS of 0.02. If we examine the Akaike information criterion (AIC), a widely-used metric of in-sample fit, a different story emerges. For both agreement failure and duration, the model with differential item functioning has the best (lowest) AIC. Moving to out-of-sample accuracy reveals a very different story. The differential item functioning model performs the best at predicting agreement failure with an AUC of 0.71. When predicting agreement duration, the full, conflict level information, robust, and differential item functioning models tie with an IBS of 0.03. These patterns suggest that the higher accuracy of other models in Table 2 may actually represent overfitting when difficulty and discrimination parameters are not allowed to vary by conflict incompatibility.

Comparing these out-of-sample results with those from agreement strength and an additive index further

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Agreement Type	0.36*						
	(0.12)						
Additive Index		−0.05*					
		(0.02)					
Baseline model			−0.28*				
			(0.10)				
Full model				−0.30*			
				(0.10)			
Conflict-level model					−0.29*		
					(0.10)		
Robust model						−0.30*	
						(0.10)	
Differential model							−0.35*
							(0.10)
In-sample IBS	0.03	0.02	0.03	0.04	0.03	0.04	0.05
3-fold IBS	0.21	0.09	0.07	0.03	0.03	0.03	0.03
AIC	1663.97	1668.22	1664.28	1663.00	1663.62	1663.33	1659.93
Num. events	152	152	152	152	152	152	152
Num. obs.	328	328	328	328	328	328	328
PH test	0.07	0.85	0.58	0.93	0.99	0.96	0.60

*p < 0.05

Table 4: Cox proportional hazard regression models of agreement strength and peace survival

illustrates the benefits of the increased nuance introduced by differential item functioning. Across both outcomes, the additive index performs worse than all model variations in out-of-sample accuracy with agreement type performing worse still. An additive index for agreement failure has an AUC of 0.40, which is 44% worse than the differential item functioning model's AUC of 0.71. Similarly, the additive index yields an IBS of 0.03 for agreement duration, 67% worse than all models except the baseline one. The model with differential item functioning is thus the one to use in scenarios that require a general measure of agreement strength at the time of signing.

4.2 Internal validity

I next move on to assessing the internal validity of the differential item functioning model by exploring whether the estimates comport with our substantive understanding of conflict resolution. Figure 2 presents the posterior means and 95% credible intervals for the difficulty parameters $\boldsymbol{\alpha}$ and the discrimination parameters $\boldsymbol{\gamma}$. Higher (lower) difficulty parameters indicate provisions that are more (less) likely to appear in stronger (weaker) agreements. Difficulty parameters define a 'baseline' of strength that agreements must

surpass for there to be a reasonable chance of observing that provision.

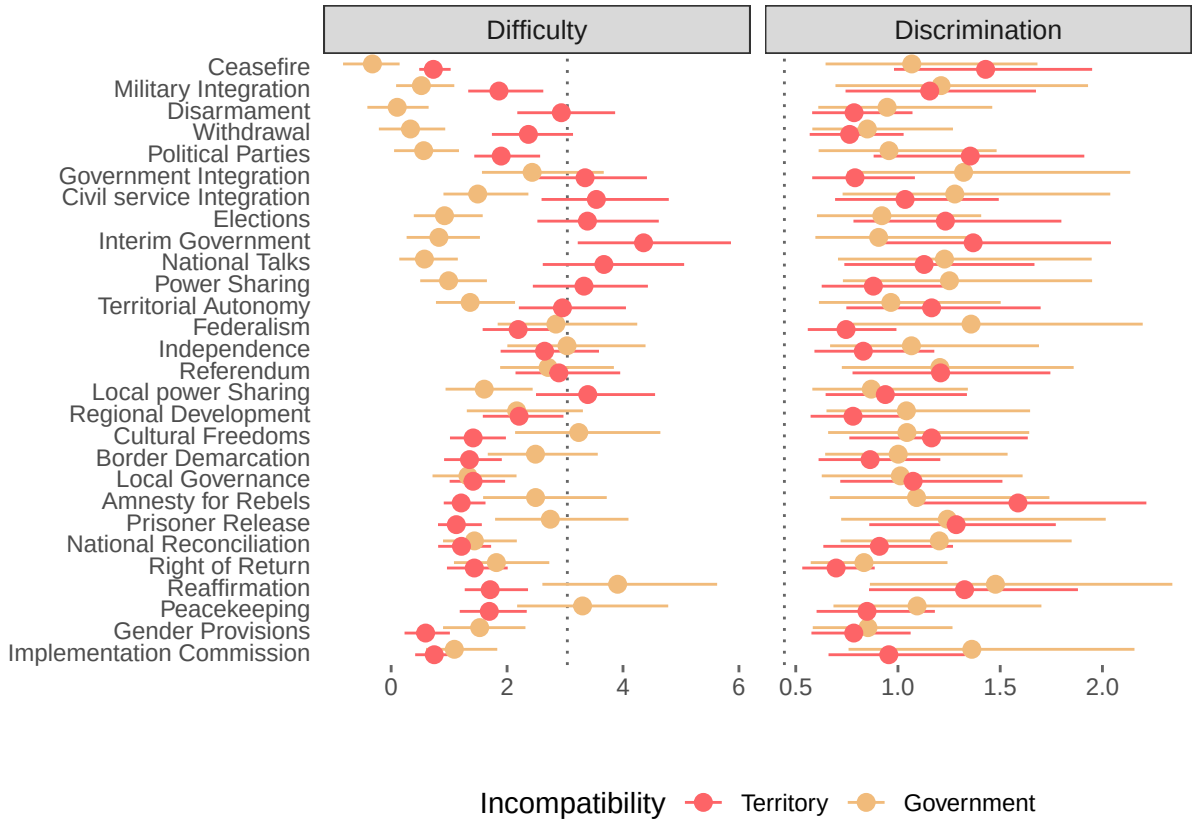

Figure 2: Posterior means and 95% credible intervals with prior means denoted by dotted lines

Power sharing provisions have an estimated α value of 0.88 in conflicts over government and 3.41 in conflicts over territory. The estimates for γ are 1.46 and 0.86, respectively. Agreements in governmental conflicts need to be less strong than those in territorial conflicts in order to have power sharing provisions, which makes sense because territorial conflicts are more likely to end with independence or autonomy for rebel groups. The much higher γ estimate in governmental conflicts means that an agreement signed in a conflict over government will likely be extremely weak if it does not have power sharing provisions. Many of the provisions in the upper half of Figure 2—such as military integration, disarmament, and the formation of an interim government—are much more relevant to governmental conflicts than territorial ones. They are thus more frequently observed in agreements in these types of conflict, and their absence indicates a particularly weak agreement. Power sharing arrangements are a costly signal that indicates that governments are serious about securing peace (Jarstad and Nilsson 2008; Martin 2013), which aligns with them having one of the higher discrimination parameter estimates in governmental conflicts.

Just as the identification restriction on β_{type} appears justified, constraining $\gamma > 0$ is defensible because none of the credible intervals for γ in the right panel of Figure 2 approach 0. Figure 2 also provides evidence that the model draws on patterns of co-occurrence and does not simply treat less common provisions as indicative of stronger agreements. Despite disarmament provisions featuring in twice as many territorial conflict agreements than military integration provisions (21 compared to 9), they have a higher estimated difficulty parameter (2.94) than military integration provisions (1.86). If the model just captured the frequencies of various provisions independent of one another, the difficulty parameter for military integration provisions would be lower than for disarmament ones because they appear in more agreements. This pattern also occurs in conflicts over government where rebel amnesty provisions appear more frequently (59) than civil service integration ones (30) despite having a higher difficulty parameter (2.49 compared to 1.49). The fact that a more common provision can have a substantially higher difficulty parameter estimate indicates that the model is not simply using frequency as a way to judge the contribution of a provision.

The relationship between observed provisions and latent strength can be made clearer by examining the item characteristic curve (ICC) for specific provisions. The ICC for provision j is simply Equation 11 evaluated across the range of $\hat{\theta}$ using $\hat{\alpha}_j$ and $\hat{\gamma}_j$.[12] Figure 3 depicts the ICCs for cultural freedoms, disarmament, and national reconciliation, along with 95% posterior uncertainty, as well as observed values of y_{ij} for the provisions.[13] Parameter estimates for governmental and territorial conflicts are represented as separate colors.

The ICC for disarmament in governmental conflicts is much steeper than in territorial conflicts because while the observed instances of disarmament ($Y_{ij} = 1$) and no disarmament ($y_{ij} = 0$) in governmental conflicts overlap, agreements with disarmament provisions are appear lower on the scale of agreement strength. In contrast, disarament provisions are spread much more widely along on scale of agreement strength in territorial conflicts, so the slope of the ICC is much lower and it does not effectively discriminate between weak and strong agreements. The greater relevance of disarmament provisions in governmental conflicts makes sense given that former belligerents must share the same territory, so a reduction in arms is necessary to reduce the incentive to renege and ensure a stable peace Walter (1997). Conversely, the ICC for

[12]For the baseline, full, or conflict-level model this would be Equation 1, and Equation 10 for the robust model.
[13]See the Supplemental Information for similar plots for all 28 provisions.

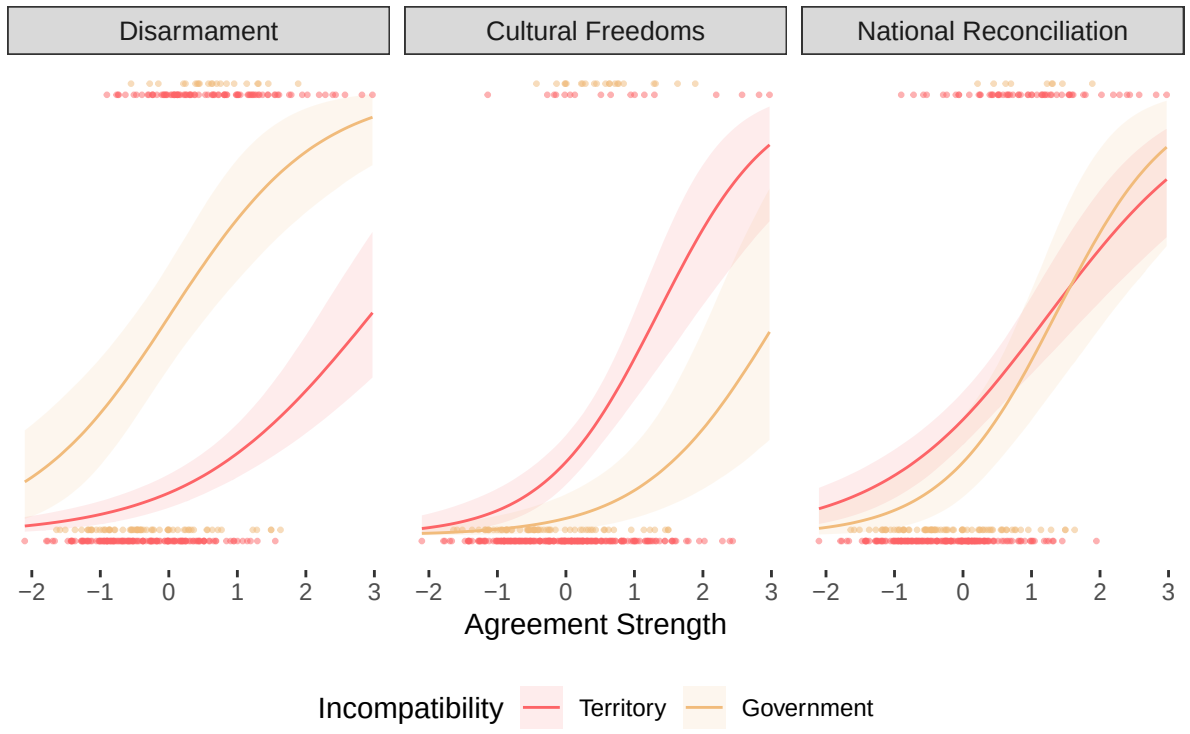

Figure 3: Item characteristic curves with differential item functioning by conflict incompatibility and observed provisions

cultural freedoms is steeper in territorial conflicts as these conflicts often have an identity-based dimension to them. The ICCs for national reconciliation in both governmental and territorial conflicts are statistically indistinguishable from one another not because national reconciliation provisions do not matter, but because they matter in both types of conflicts; the separation between agreements with and without such provisions is stronger than for disarmament or cultural freedoms provisions.

We can also qualitatively inspect the data to assess the convergent validity of the strength measure. Figure 4 presents provisions in the 10 strongest agreements, in comparison with the selected agreements in Figure 1. All 10 of these agreements have ceasefire, disarmament, government integration, elections, power sharing, national reconciliation, right of return, and implementation provisions. Given the importance of ceasefires as necessary preconditions for peace, and the many ways in which power sharing works to bind former combatants to peaceful cooperation, it is unsurprising that both of them are ubiquitous among the strongest agreements. The low difficulty and relatively high discrimination of ceasefire provisions (for both territorial and governmental conflicts) indicate that strong agreements have ceasefire provisions and weak ones do not, so if an agreement lacks a ceasefire provision we believe that it is weak, and we have high

confidence in that belief. Similarly, power sharing (Hartzell and Hoddie 2003) and elections (Matanock 2017) are mechanisms to strengthen peace that have been extensively studied, so their pervasiveness among the strongest agreements is unsurprising.

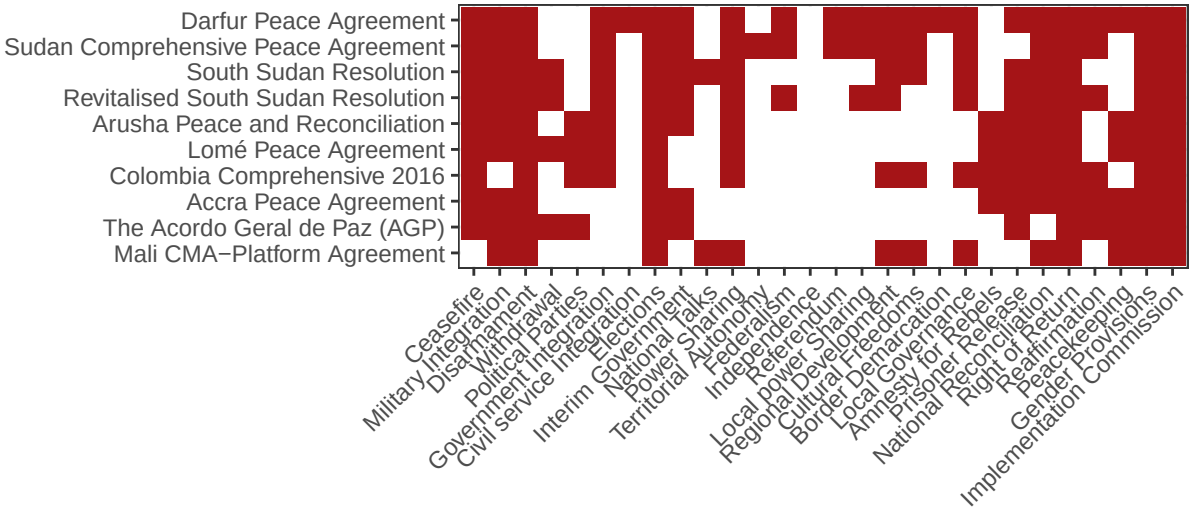

Figure 4: Provisions in ten strongest peace agreements

90% of the 10 strongest agreements are signed in territorial conflicts. While there are 2.81 times as many agreements in territorial conflicts in governmental ones, the difference in the distributions of strength between them is obscured by an additive index. The difference between the average number of provisions in governmental conflicts (5.43) and in territorial ones (5.55) as a percentage of the total range of provisions (0.54%) is smaller than this quantity for estimates from the differential model (2.34%). This difference is especially noticeable when comparing the distributions of estimates across incompatibility in Figure 5. 25 agreements have zero provisions, but only 14 of these provision-less agreements appear in the 25 weakest agreements. This result demonstrates the added value of including information from PA-X and PAM, which introduces additional variation beyond the UCDP Peace Agreements Data.

Goertz (2009, 9–11) notes that in addition to qualitatively exploring extreme cases of a measure, we must also examine the distribution of cases at the extremes of the scale. A measure with high density at either extreme would suggest a scale that continues past the measured values and consequently cast doubt on the suitability of the measurement strategy. Figure 5 illustrates a moderate degree of positive skew with a short tail of strong agreements in territorial conflicts, but not a high concentration of them, providing evidence of

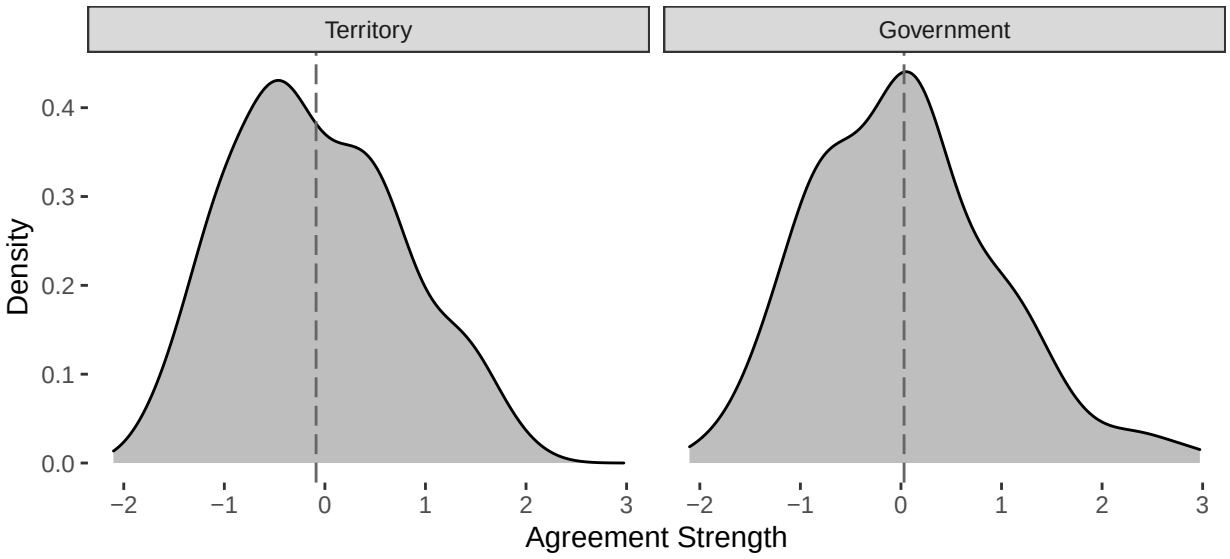

Figure 5: Distribution of agreement strength measures with mean denoted by dashed line

construct validity.[14]

5 Peace agreement strength scores

With the model validated, I now present the distribution of estimates. Figure 6 displays the posterior mean of each agreement's strength, as well as its 95% credible interval with the agreements described in Figure 1 labeled. Agreements near the top of the scale have the least uncertainty because they have the highest number of provisions, so the model has the most information on them. While there is considerable uncertainty around the estimates, many agreements are substantially different from one another as the 95% credible intervals do not overlap. Although there are 80 agreements with non-unique patterns of provisions in the UCDP Peace Agreement Data, 0 agreements have identical θ values due to the extra information contributed by PA-X and PAM.

Comparing the scores with existing measures of peace agreement strength allows us to evaluate their convergent validity; I do so by measuring the statistical association between the estimates and an additive index of agreement provisions, which has been used to capture agreement strength (Werner and Yuen 2005). As an additive index is an ordinal variable, the Spearman rank correlation coefficient is a better measure of

[14]While the prior on θ is a normal distribution, the strength of the data is sufficient to generate an asymmetric distribution of agreement strengths, indicating that the strength estimates are not solely a product of the priors.

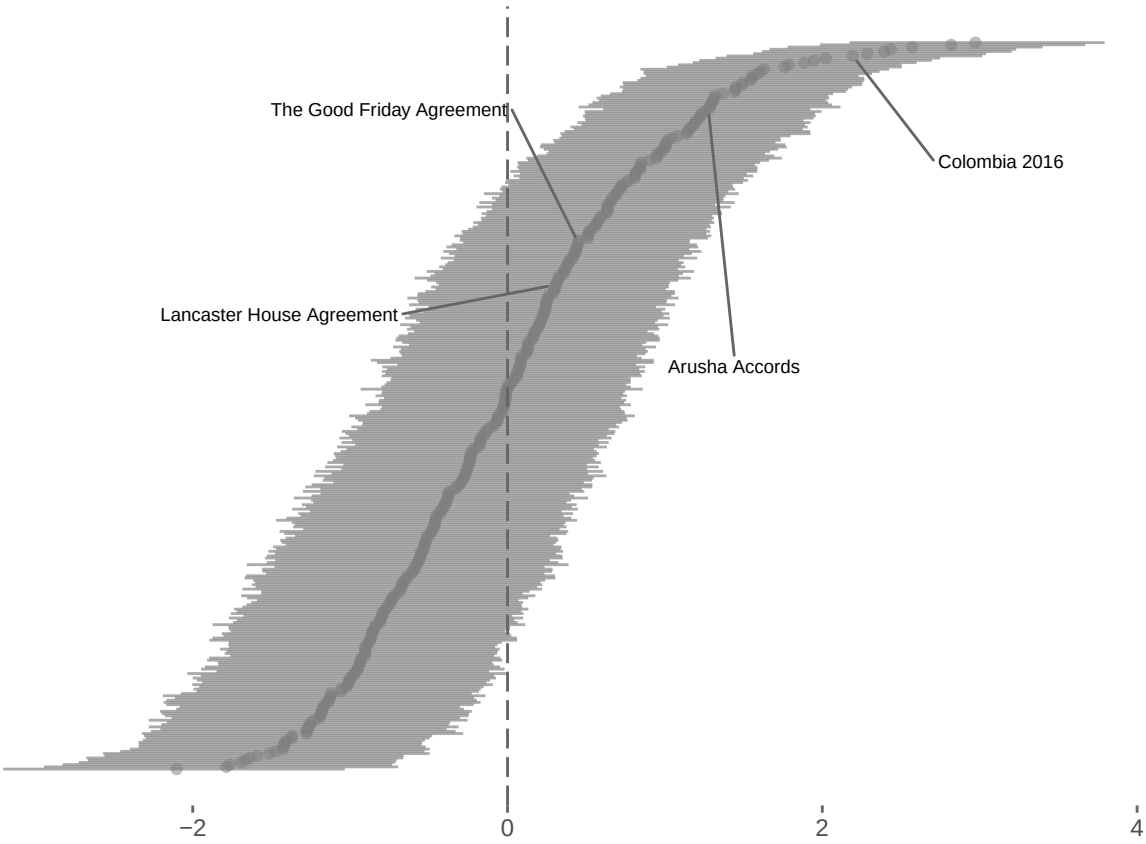

Figure 6: Distribution of agreement strengths with 95% credible intervals

association than the Pearson correlation coefficient. However, Spearman's ρ is biased in the presence of ties in one or more variables, and all but 2 values of an additive index have ties. The Kendall rank correlation coefficient $\tau_{\mathcal{B}}$ corrects for ties and $\tau_{\mathcal{B}}$ between an additive index and the estimates is 0.77. This positive but not perfect correlation suggests that the extra information contained in the measurement model introduces substantial nuance into the strength scores.

Exploring where these measures disagree is instructive. Statistical associations between different measures can mask patterns of variation between them, especially when they agree on cases at the ends of the spectrum (Goertz and Mahoney 2012, 133–36). Figure 7 plots the rank ordering of agreements under an additive agreement against the rank ordering of the estimates, rather than the estimates themselves because the units of the latent scale are not inherently meaningful.

Agreement is high at the strong and weak ends of the spectrum, but decreases sharply towards the center. Importantly, this disagreement does not represent just a decrease in statistical efficiency where the ranking of

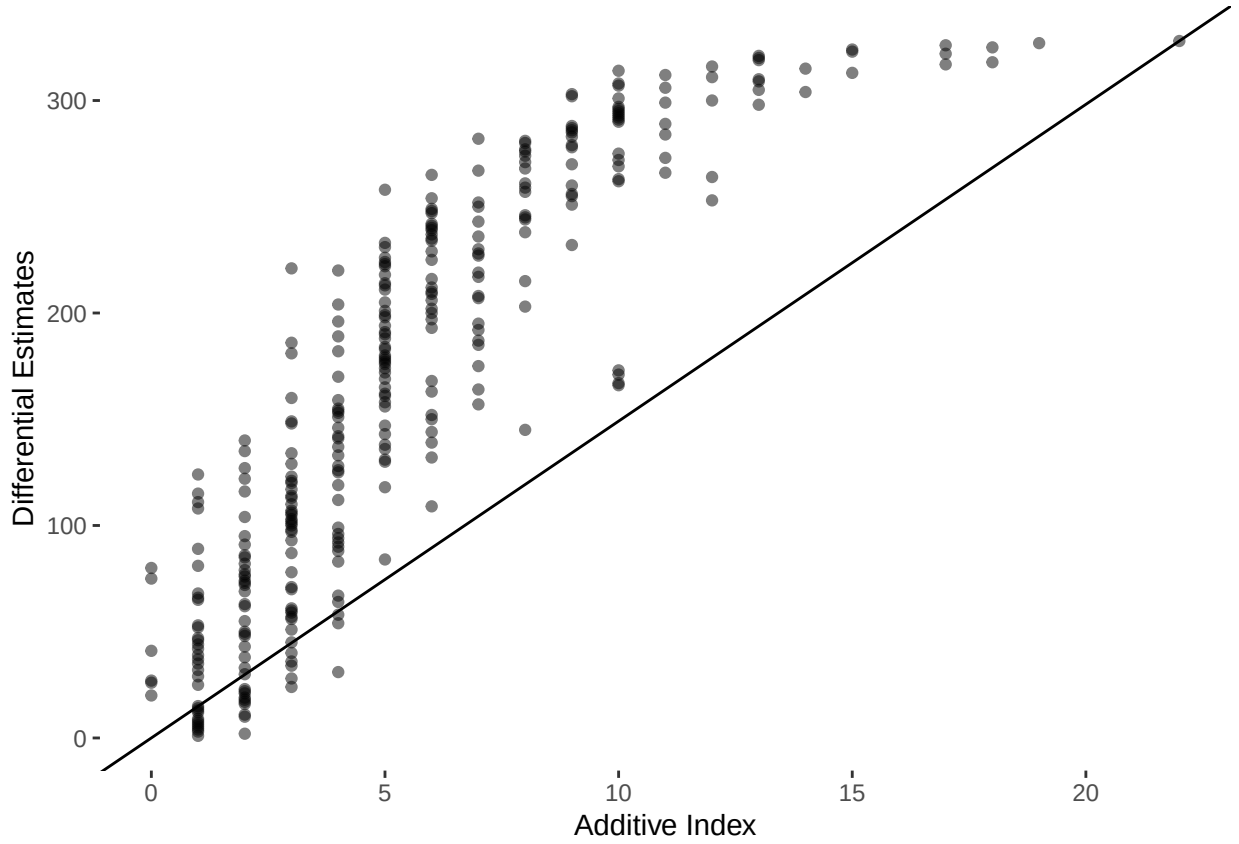

Figure 7: Rank ordering of agreement strengths measured by estimates from model with differential item functioning and an additive index

similar agreements is switched. If this were the case, each cluster of points would be centered around the 45° line. Instead, we see that an additive index systematically underestimates the strength of agreements, most extremely in the middle of the spectrum. This variation may explain the higher out-of-sample predictive accuracy of the estimates compared to an additive index in Tables 3 and 4.

6 Selection effects in agreement signing

While Section 4 is focused on comparing the external predictive validity of the various measurement models, the substantive implications of the results in Tables 3 and 4 are counterintuitive. Strong agreements appear to be associated with an increased risk of conflict recurrence and decreased survival time until recurrence. One possibility is that there is a selection effect where stronger agreements are signed at the conclusion of more intractable conflicts. The factors that make these conflicts more difficult to resolve could then continue

to affect the post-conflict peace, reducing its likelihood of survival. A direct measure of peace agreement strength allows us to investigate this potential confounding effect on the relationship between agreement content and duration.

Strong selection effects characterize the presence or absence of specific provisions within peace agreements. Hartzell and Hoddie (2015) find that conflicts with higher levels of distrust among former combatants — those that lasted longer, were recurrent conflicts, or occurred in highly fractionalized societies — are more likely to generate peace agreements with multiple power sharing provisions. Similarly, Fortna (2008) finds that UN peacekeeping missions are most likely to be dispatched in the most difficult to resolve conflicts and Binningsbø and Rustad (2012) find that economic power sharing agreements are most likely to be implemented following conflicts fought over natural resources. This evidence suggests that a selection effect on agreement strength and conflict characteristics is plausible.

As an initial evaluation of the possibility that the strongest agreements are likewise signed in the toughest conflicts, I evaluate the relationship between multiple indicators of conflict intractability and agreement strength. Conflicts with many combatants take longer to resolve due to the increased number of veto players (Cunningham 2006), so civil wars that are coded as *internationalized* in the ACD may also have stronger agreements due to the need to satisfy more stakeholders (Pettersson, Högbladh, and Öberg 2019). Conflicts fought over identity can be some of the most difficult to resolve (Licklider 1995; Denny and Walter 2014), so I use the Ethnic Power Relations (EPR) dataset to code whether a conflict is an *ethnic conflict* (Vogt et al. 2015). I fit models using both potential confounders to explain estimates of agreement strength from the differential item functioning measurement model. I also estimate models that control for whether an agreement was signed during the *Cold War* and *conflict duration*, which would not be possible if this conflict-level information was included in the prior on θ. Table 5 presents results from these models.

The *internationalized* and *ethnic conflict* variables are positive and statistically significant in all specifications, suggesting that stronger agreements are more likely in these less easily resolved conflicts. Ending during the *Cold War* is negatively and significantly associated with weaker agreements, which aligns with the finding by Hartzell & Hoddie that conflicts ending after the *Cold War* are more likely to implement multiple forms of power sharing (2015). Taken together these findings provide evidence that harder to resolve conflicts are

	Model 1	Model 2	Model 3	Model 4
Internationalized	0.39*	0.32*		
	(0.14)	(0.13)		
Ethnic Conflict			0.59*	0.49*
			(0.14)	(0.13)
Cold War		−0.63*		−0.56*
		(0.16)		(0.16)
Conflict Duration		−0.01*		−0.01*
		(0.00)		(0.00)
(Intercept)	−0.06	0.22*	−0.50*	−0.16
	(0.05)	(0.09)	(0.12)	(0.15)
Adj. R^2	0.02	0.07	0.05	0.10
Num. obs.	328	328	328	328

*p < 0.05

Table 5: Linear models of conflict severity and agreement strength

associated with stronger agreements.

7 Conclusion

These scores allow researchers to comprehensively measure the strength of a peace agreement at the time of signing, in contrast to previous approaches that either rely on agreement duration as a rough proxy of strength, or employ only a handful of conflict resolution provisions to measure agreement strength. While the measure developed by Williams et al. (2021) resolves both of these issues, it depends on conflict-level information such as the presence of third-party mediators. This dependence precludes it from being used to analyze how the process by which an agreement is reached can affect its strength (Albin and Druckman 2012; Druckman and Wagner 2019), which the estimates can do.

In contrast, the measurement model is able to identify a selection effect in the design of peace agreements. While some studies of peace agreement durability such as Hartzell and Hoddie (2015) account for this selection effect, many do not. This selection effect suggests that future studies of agreement durability should control for not only the strength of an agreement at time of signing, but also the process that affected the content of the agreement.

The estimates are negatively correlated with agreement duration ($\rho = -0.28$ among agreements that ended in failure), and a Cox proportional-hazards regression finds a negative and statistically significant relationship

between agreement strength and duration.[15] One possible explanation is that stronger agreements are associated with a lower degree of ultimate implementation due to backlash from newly disadvantaged stakeholders. This would mean that any relationship between agreement strength at the time of signing and the ultimate fate of that agreement would be mediated by the intervening degree of implementation. Another alternative is that the more wide-ranging provisions associated with stronger agreements are more difficult to implement, and the failure of implementation could generate new grievances severe enough to overcome collective action problems and incite new conflicts in the future.

In order to adjudicate between these possibilities, more research is needed to disentangle the relationship between *de jure* institutions embodied in agreement provisions and *de facto* practices after the conflict. In doing so, we will gain a better understanding of how both affect post-conflict outcomes. Because the measurement model relies only on the actual content of peace agreements, it is ideally-suited to this goal. By using only information contained within agreements themselves, the model provides a way to capture the independent effect of post-conflict institutions. Exploring these dynamics will likely shed light on the specific causal mechanisms by which post-conflict societies remain stable or return to widespread political violence.

One downside to the model and the estimates it produces is that it cannot account for the intent behind a provision's inclusion in an agreement. The data do not say whether a provision was a must-have protection demanded by one or more of the belligerents, or was introduced by a third-party mediator and met with indifference by the parties. The former case should have a much larger impact on the strength of an agreement, but the model treats both equally. Future data collection efforts should expand on existing data sources by investigating and recording the processes that led to the signing of individual agreements. While this may not be possible in all cases, especially with small early-stage agreements, many conflict resolution efforts involve multiple outside mediators and are highly watched affairs.

Another limitation of the model is that it only measures the strength of peace agreements signed after civil conflicts. Many of the institutions laid out in negotiated settlements can be implemented after the cessation of hostilities where conflict ended in either outright victory for one side or a prolonged period of low activity on the part of the rebels. In either of these cases, the parties are unlikely to reach and ratify a formal agreement.

[15] See SI for details.

This means that the model cannot address larger questions about the role of post-conflict institutions more broadly than those established as part of negotiated settlements. However, this shortcoming highlights a way forward in the study of peace agreements after civil war.

If scholars gather data on the presence or absence of the institutions stipulated in peace agreements in all post-conflict societies, then we will be able to answer whether including provisions for them in negotiated settlements increases their likelihood of being implemented. More importantly, this will assist scholars in isolating the independent effect of peace agreements on post-conflict stability. If stronger peace agreements are associated with a lower risk of renewed violence, even accounting for the implementation or not of specific provisions in all post-conflict societies, then that would suggest that the perceived effect of peace agreements is not endogenous. In the same manner that Gates et al. (2016) measure different forms of power sharing institutions in all societies—post-conflict or not—scholars should collect data on the implementation of the institutions outlined in peace agreements in the years following all civil conflicts, regardless of whether the conflict ended in a negotiated settlement, outright victory by one combatant, or a period of low activity. Doing so will allow us to isolate the independent effect of peace agreements on post-conflict stability, contributing rigor to the larger debate over the effect of institutions, which would not be possible without a robust measure of agreement strength at the time of signing.

References

Albin, Cecilia, and Daniel Druckman. 2012. "Equality Matters: Negotiating an End to Civil Wars." *Journal of Conflict Resolution* 56 (2): 155–82. https://doi.org/10.1177/0022002711431798.

Bafumi, Joseph, Andrew Gelman, David K. Park, and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13 (2): 171–87. https://doi.org/10.1093/pan/mpi010.

Bell, Christine, and Sanja Badanjak. 2019. "Introducing PA-X: A New Peace Agreement Database and Dataset." *Journal of Peace Research* 56 (3): 452–66. https://doi.org/10.1177/0022343318819123.

Binningsbø, Helga Malmin, and Siri Aas Rustad. 2012. "Sharing the Wealth: A Pathway to Peace or a Trail to Nowhere?" *Conflict Management and Peace Science* 29 (5): 547–66. https://doi.org/10.1177/0738894212456952.

Bradley, Andrew P. 1997. "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recognition* 30 (7): 1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.

Buhaug, Halvard. 2006. "Relative Capability and Rebel Objective in Civil War." *Journal of Peace Research* 43 (6): 691–708. https://doi.org/10.1177/0022343306069255.

Carter, Jeff, and Charles E. Smith Jr. 2020. "A Framework for Measuring Leaders' Willingness to Use Force." *American Political Science Review* 114 (4): 1352–58. https://doi.org/10.1017/S0003055420000313.

Cederman, Lars-Erik, Halvard Buhaug, and Jan Ketil Rød. 2009. "Ethno-Nationalist Dyads and Civil War A GIS-Based Analysis." *Journal of Conflict Resolution* 53 (4): 496–525. https://doi.org/10.1177/0022002709336455.

Cederman, Lars-Erik, Simon Hug, Andreas Schädel, and Julian Wucherpfennig. 2015. "Territorial Autonomy in the Shadow of Conflict: Too Little, Too Late?" *American Political Science Review* 109 (2): 354–70. https://doi.org/10.1017/S0003055415000118.

Cunningham, David E. 2006. "Veto Players and Civil War Duration." *American Journal of Political Science* 50 (4): 875–92.

Denny, Elaine K., and Barbara F. Walter. 2014. "Ethnicity and Civil War." *Journal of Peace Research* 51 (2): 199–212. https://doi.org/10.1177/0022343313512853.

Druckman, Daniel, and Lynn Wagner. 2019. "Justice Matters: Peace Negotiations, Stable Agreements, and Durable Peace." *Journal of Conflict Resolution* 63 (2): 287–316. https://doi.org/10.1177/0022002717739088.

Efron, Bradley. 1983. "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation." *Journal of the American Statistical Association* 78 (382): 316–31. https://doi.org/10.1080/01621459.1983.10477973.

Fortna, Virginia. 2003. "Scraps of Paper? Agreements and the Durability of Peace." *International Organization* 57 (2): 337–72.

———. 2008. *Does Peacekeeping Work? Shaping Belligerents' Choices After Civil War*. Princeton: Princeton University Press.

Gates, Scott, Benjamin A. T. Graham, Yonatan Lupu, Håvard Strand, and Kaare W. Strøm. 2016. "Power Sharing, Protection, and Peace." *The Journal of Politics* 78 (2): 512–26. https://doi.org/10.1086/684366.

Goertz, Gary. 2009. "Concepts, Theories, and Numbers: A Checklist for Constructing, Evaluating, and Using Concepts or Quantitative Measures." In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, 1–29. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286546.003.0005.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures : Qualitative and Quantitative Research in the Social Sciences*. Princeton, N.J.: Princeton University Press.

Graf, Erika, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. "Assessment and Comparison of Prognostic Classification Schemes for Survival Data." *Statistics in Medicine* 18 (17-18): 2529–45. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5.

Harbom, Lotta, Stina Högbladh, and Peter Wallensteen. 2006. "Armed Conflict and Peace Agreements." *Journal of Peace Research* 43 (5): 617–31. https://doi.org/10.1177/0022343306067613.

Hartzell, Caroline. 1999. "Explaining the Stability of Negotiated Settlements to Intrastate Wars." *Journal of Conflict Resolution* 43 (1): 3–22. https://doi.org/10.1177/0022002799043001001.

Hartzell, Caroline, and Matthew Hoddie. 2003. "Institutionalizing Peace: Power Sharing and Post-Civil War Conflict Management." *American Journal of Political Science* 47 (2): 318–32. https://doi.org/10.1111/1540-5907.00022.

———. 2015. "The Art of the Possible: Power Sharing and PostCivil War Democracy." *World Politics* 67 (1): 37–71.

———. 2019. "Power Sharing and the Rule of Law in the Aftermath of Civil War." *International Studies Quarterly* 63 (3): 641–53. https://doi.org/10.1093/isq/sqz023.

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

Hastie, Trevor, Jerome H. Friedman, and Robert Tibshirani. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition, corrected 12th printing. New York: Springer.

Hathaway, Oona A. 2002. "Do Human Rights Treaties Make a Difference?" *The Yale Law Journal* 111 (8): 1935–2042. https://doi.org/10.2307/797642.

Howard, Lise Morjé, and Alexandra Stark. 2018. "How Civil Wars End: The International System, Norms, and the Role of External Actors." *International Security* 42 (3): 127–71. https://doi.org/10.1162/ISEC_a_00305.

Jarstad, Anna K., and Desiree Nilsson. 2008. "From Words to Deeds: The Implementation of Power-Sharing Pacts in Peace Accords." *Conflict Management and Peace Science* 25 (3): 206–23. https://doi.org/10.1080/07388940802218945.

Johnson, Valen E., and James H. Albert. 1999. *Ordinal Data Modeling*. Statistics for Social and Behavioral Sciences. New York: Springer-Verlag. https://doi.org/10.1007/b98832.

Joshi, Madhav, and J. Michael Quinn. 2015. "Is the Sum Greater Than the Parts? The Terms of Civil War Peace Agreements and the Commitment Problem Revisited: Civil War Peace Agreements." *Negotiation Journal* 31 (1): 7–30. https://doi.org/10.1111/nejo.12077.

Joshi, Madhav, and Jason Michael Quinn. 2017. "Implementing the Peace: The Aggregate Implementation of

Comprehensive Peace Agreements and Peace Duration After Intrastate Armed Conflict." *British Journal of Political Science* 47 (4): 869–92. https://doi.org/10.1017/S0007123415000381.

Joshi, Madhav, Jason Michael Quinn, and Patrick M Regan. 2015. "Annualized Implementation Data on Comprehensive Intrastate Peace Accords, 1989." *Journal of Peace Research* 52 (4): 551–62. https://doi.org/10.1177/0022343314567486.

Keohane, Robert O. 1988. "International Institutions: Two Approaches." *International Studies Quarterly* 32 (4): 379–96. https://doi.org/10.2307/2600589.

Licklider, Roy. 1995. "The Consequences of Negotiated Settlements in Civil Wars, 1945-1993." *The American Political Science Review* 89 (3): 681–90. https://doi.org/10.2307/2082982.

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

Lo, Nigel, Barry Hashimoto, and Dan Reiter. 2008. "Ensuring Peace: Foreign-Imposed Regime Change and Postwar Peace Duration, 1914." *International Organization* 62 (4): 717–36. https://doi.org/10.1017/S0020818308080259.

Martin, Philip. 2013. "Coming Together: Power-Sharing and the Durability of Negotiated Peace Settlements." *Civil Wars* 15 (3): 332–58. https://doi.org/10.1080/13698249.2013.842747.

Matanock, Aila M. 2017. "Bullets for Ballots: Electoral Participation Provisions and Enduring Peace After Civil Conflict." *International Security* 41 (4): 93–132.

Mattes, Michaela, and Burcu Savun. 2009. "Fostering Peace After Civil War: Commitment Problems and Agreement Design." *International Studies Quarterly* 53 (3): 737–59. https://doi.org/10.1111/j.1468-2478.2009.00554.x.

———. 2010. "Information, Agreement Design, and the Durability of Civil War Settlements." *American Journal of Political Science* 54 (2): 511–24. https://doi.org/10.1111/j.1540-5907.2010.00444.x.

Ottmann, Martin, and Johannes Vüllers. 2015. "The Power-Sharing Event Dataset (PSED): A New Dataset on the Promises and Practices of Power-Sharing in Post-Conflict Countries." *Conflict Management and Peace Science* 32 (3): 327–50. https://doi.org/10.1177/0738894214542753.

Pettersson, Therése, Stina Högbladh, and Magnus Öberg. 2019. "Organized Violence, 1989-2018

and Peace Agreements." *Journal of Peace Research*, June, 1–15. https://doi.org/10.1177/0022343319856046.

Pillar, Paul. 1983. *Negotiating Peace: War Termination as a Bargaining Process*. Princeton, N.J.: Princeton University Press.

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

Reid, Lindsay. 2017. "Finding a Peace That Lasts: Mediator Leverage and the Durable Resolution of Civil Wars." *Journal of Conflict Resolution* 61 (7): 1401–31. https://doi.org/10.1177/0022002715611231.

Ruggeri, Andrea, Han Dorussen, and Theodora-Ismene Gizelis. 2017/ed. "Winning the Peace Locally: UN Peacekeeping and Local Conflict." *International Organization* 71 (1): 163–85. https://doi.org/10.1017/S0020818316000333.

Rustad, Siri Aas, and Helga Malmin Binningsbø. 2012. "A Price Worth Fighting for? Natural Resources and Conflict Recurrence." *Journal of Peace Research* 49 (4): 531–46. https://doi.org/10.1177/0022343312444942.

Simmons, Beth A. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs." *American Political Science Review* 94 (4): 819–35. https://doi.org/10.2307/2586210.

———. 2009. *Mobilizing for Human Rights: International Law in Domestic Politics*. Cambridge: Cambridge Univ. Press.

Simmons, Beth A., and Daniel J. Hopkins. 2005. "The Constraining Power of International Treaties: Theory and Methods." *American Political Science Review* 99 (4): 623–31. https://doi.org/10.1017/S0003055405051920.

Svensson, Isak. 2007. "Mediation with Muscles or Minds? Exploring Power Mediators and Pure Mediators in Civil Wars." *International Negotiation* 12 (2): 229–48. https://doi.org/10.1163/138234007X223294.

———. 2009. "Who Brings Which Peace? Neutral Versus Biased Mediation and Institutional Peace Arrangements in Civil Wars." *Journal of Conflict Resolution* 53 (3): 446–69. https://doi.org/10.

1177/0022002709332207.

Toft, Monica. 2003. *The Geography of Ethnic Violence : Identity, Interests, and the Indivisibility of Territory*. Princeton, N.J.: Princeton University Press.

Toft, Monica Duffy. 2014. "Territory and War." *Journal of Peace Research* 51 (2): 185–98. https://doi.org/10.1177/0022343313515695.

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin. 2015. "Integrating Data on Ethnicity, Geography, and Conflict The Ethnic Power Relations Data Set Family." *Journal of Conflict Resolution* 59 (7): 1327–42. https://doi.org/10.1177/0022002715591215.

von Stein, Jana. 2005. "Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance." *American Political Science Review* 99 (4): 611–22. https://doi.org/10.1017/S0003055405051919.

Walter, Barbara. 1997. "The Critical Barrier to Civil War Settlement." *International Organization* 51 (03): 335–64. https://doi.org/10.1162/002081897550384.

———. 2002. *Committing to Peace: The Successful Settlement of Civil Wars*. Princeton, N.J.: Princeton University Press.

Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke. 2010. "The Perils of Policy by p-Value: Predicting Civil Conflicts." *Journal of Peace Research* 47 (4): 363–75. https://doi.org/10.1177/0022343309356491.

Werner, Suzanne. 1999. "The Precarious Nature of Peace: Resolving the Issues, Enforcing the Settlement, and Renegotiating the Terms." *American Journal of Political Science* 43 (3): 912–34. https://doi.org/10.2307/2991840.

Werner, Suzanne, and Amy Yuen. 2005. "Making and Keeping Peace." *International Organization* 59 (2): 261–92. https://doi.org/10.1017/S0020818305050095.

Williams, Rob, Daniel J. Gustafson, Stephen E. Gent, and Mark J. C. Crescenzi. 2021. "A Latent Variable Approach to Measuring and Explaining Peace Agreement Strength." *Political Science Research and Methods* 9 (1): 89–105. https://doi.org/10.1017/psrm.2019.23.