PASS: Peace Agreement Strength Scores*

Rob Williams†

January 02, 2020

Abstract

This paper presents the Peace Agreement Strength Scores (PASS) of peace agreements in civil conflicts. The scores capture the strength of peace agreements at the time of signing and can be used to avoid relying on the duration of agreement survival as a proxy for agreement strength. The scores are used to 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. The scores are available for all peace agreements signed in UCDP/PRIO civil conflicts from 1975-2018.

Word count: 5,848

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. This manuscript introduces the Peace Agreement Strength Scores, PASS, a general measure of agreement strength for agreements signed in civil conflicts from 1975-2018. This measure allows scholars to investigate the relationship between a multitude of factors and agreement strength, as well as control for the overall strength of agreements when studying the relationship between specific provisions and post-conflict outcomes.

PASS offers multiple advantages over existing measures of peace agreement strength. Relying on the duration of agreements to capture their strength at the time of signing risks introducing post-treatment measurement error due to the influence of post-signing factors on their duration. Picking one or two

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

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. Finally, Williams et al. (n.d.) employ a similar latent variable model but rely on conflict-level information that prevent their measure from being employed in any analysis involving the same conflict-level variables. PASS avoids all of these shortcomings, resulting in a nuanced and broadly-applicable measure of peace agreement strength.

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 (Walter 2002) due to the need to integrate former combatants into society (Hartzell 1999; Hartzell, Hoddie, and Rothchild 2001; Hartzell and Hoddie 2003). For these reasons, it is not appropriate to pool peace agreement across types of conflict. The data contain only 31 peace agreements signed in international conflict due to the rarity of interstate conflict during the sample period, so PASS only measures agreements in civil conflict.

PASS relies on "conflict resolution provisions" which are "stipulations in peace agreements aiming to resolve the basic incompatibilities" (Svensson 2007, 241). These 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). If 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. Some provisions such as initiation of peacekeeping missions are designed to address commitment problems that arise in a post-conflict environment (Walter 2009), and while they do not address the root of the conflict, they should have a similarly pacifying effect on the likelihood of future conflict.

The latent variable model PASS employs 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.

1 The data

PASS employs the UCDP Peace Agreements Dataset version 19.1 (Pettersson, Högbladh, and Öberg 2019), which contains information on the provisions contained in 324 unique peace agreements from 1975-2018. Table 1 presents all 28 provisions in the data.[1]

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

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

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

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

[1]The outlining provision is omitted following exploratory analyses. See the Supplemental Information for full details.

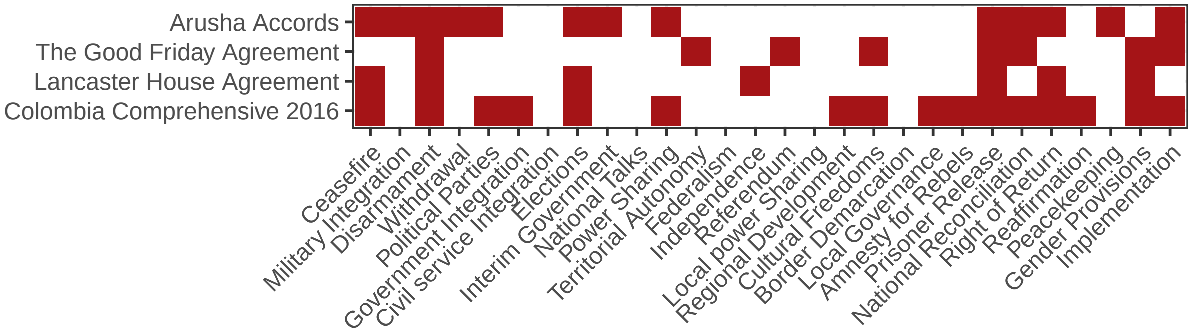

Figure 1: Provisions in various peace agreements.

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.

The data contain 3 agreements that are signed in more than one conflict. Table 2 presents these agreements.

Agreement	State	Year	Conflicts
Vance-Owen Plan	Bosnia-Herzegovina	1993	2
Nationwide Ceasefire	Myanmar (Burma)	2015	3
Deed of Commitment	Myanmar (Burma)	2015	2

Table 2: Multiple conflict agreements

I deal with these cases by splitting the agreements, creating one observation per agreement-conflict pair because the same agreement may be stronger or weaker in different conflicts due to different underlying issues driving the violence. An agreement with the same provisions may be stronger in one conflict because it addresses more of the rebels' grievances and weaker in another because of a mismatch between the provisions and the second group's grievances. This approach makes conceptual sense because an agreement signed between a government and multiple rebel groups in multiple conflicts does not automatically fail when one conflict restarts. While the violence introduced by the recurrence of one conflict may destabilize relationships between the state and other signatories, there is no systematic evidence that the resumption of hostilities between two signatories to a multiparty agreement will undermine the peace between the other signatories (Nilsson 2008). Disaggregating these multi-conflict agreements allows for the fact that one signatory may return to violence and invalidate the deal while another continues to abide by it.

2 Measurement strategy

The strength of peace agreements is measured as a latent variable using the three parameter item response model, often called the Rasch 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:

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

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 sets the value of θ_i above which there is a non-negligible probability of observing $y_{ij} = 1$. 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 $y_{ij} = 0$.

The provisions in the data are less 'clean' than the roll calls typically found in voting data, and there are few separating hyperplanes that can divide weak and strong agreements. The 'error' parameter $\boldsymbol{\epsilon}$ accounts for this lack of separability and is traditionally used to model guessing on exams in the educational context (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.

Splitting the multi-conflict agreements identified in Section 1 introduces 3 sets of agreements with identical provisions. The model will give each set of disaggregated agreements identical $\boldsymbol{\theta}$ values as the data used to estimate them will be identical. To resolve these ties, the model uses conflict-level information to inform the prior on $\boldsymbol{\theta}$ via a random intercept, $\boldsymbol{\delta}$, for each of the 63 conflicts in the data.

The Bosnia-Herzegovina: Croat conflict contains 4 agreements and the Bosnia-Herzegovina: Serb conflict contains 5 agreements. By allowing for information sharing between θ_j values in the same conflicts, the model leverages the different provisions present in the other agreements signed in each conflict and assigns different θ_j estimates to both instances of The Vance-Owen Plan. This same process occurs for The Deed of Commitment for Peace and National Reconciliation in the Bosnia-Herzegovina: Croat, Myanmar (Burma): Shan, Myanmar (Burma): Shan, Myanmar (Burma): Government conflict as well as the Nationwide Ceasefire Agreement (NCA) in the Myanmar (Burma): Shan and Myanmar (Burma): Government conflicts, as well as the remaining 60 conflicts in the data. While this information sharing is beneficial for other agreements, it is especially important to avoid identical $\boldsymbol{\theta}$ estimates for these 3 agreements.

In addition to the conflict random intercept $\boldsymbol{\delta}$, the prior mean of $\boldsymbol{\theta}$ is defined as a linear combination of agreement-level covariates. The first is *agreement type* which 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. Further agreement-level covariates come from the PA-X dataset (Bell and Badanjak 2019), which includes much more detailed data at more disaggregated levels than the UCDP Peace Agreements Data, but only for 82.32 of the agreements in the UCDP Peace Agreements Data. Missingness is higher than the conventionally accepted threshold of 15% and PA-X excludes agreements before 1990, meaning the data are missing not at random, so imputation before including the PA-X provisions in Equation 1 is inappropriate (Little and Rubin 2002). Instead, I follow Carter and Smith Jr. (2019) and allow the PA-X provisions to inform the prior on $\boldsymbol{\theta}$ alongside *agreement type*. Table 3 presents the 6 PA-X provisions.

Civil Society	Protection Measures	Transitional Justice Mechanism
Human Rights/Rule of Law	Criminal Justice/emergency law	Reparations

Table 3: Agreement-level covariates from PA-X

These provisions were chosen because they are theoretically relevant to peace agreement strength without duplicating provisions already included in the UCDP Peace Agreements Data. While it is possible to include additional conflict-level covariates such as *incompatibility* or whether *ethnic* cleavages motivate the violence, 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. The full model with priors and hyperpriors is presented below:

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\delta} + \mathbf{X}\boldsymbol{\beta}, 1) \tag{2}$$
$$\boldsymbol{\gamma} \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma) \tag{3}$$
$$\boldsymbol{\alpha} \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha) \tag{4}$$
$$\boldsymbol{\beta} \sim \mathcal{N}(0, 1) \tag{5}$$
$$\boldsymbol{\delta} \sim \mathcal{N}(\mu_\delta, \sigma_\delta) \tag{6}$$
$$\mu_\gamma, \mu_\alpha, \mu_\delta \sim \mathcal{N}(0, 1) \tag{7}$$
$$\sigma_\alpha, \sigma_\delta \sim \exp(1) \tag{8}$$
$$\epsilon \sim \mathcal{U}(0, 0.1) \tag{9}$$

The prior on ϵ 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 agreement-level covariates are included in Equation 2 as $\mathbf{X}$. The conflict random intercept $\boldsymbol{\delta}$ accounts for unobserved heterogeneity in the data, meaning that two agreements with identical provisions will yield different strength ratings. This is actually desirable because the model is not comprehensive and necessarily fails to include some relevant information. The sharing of information across agreements in the same conflict via δ helps mitigate some of this omitted variable bias.

This model is similar to the one in Williams et al. (n.d.), 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 external factors in civil war, making their measure unsuitable for use in analyses where those factors are of interest. Neither of these limitations apply to PASS, as it does not include any conflict-level information and employs a much more general identification strategy detailed below.

Williams et al. (n.d.) find that their measure is positively correlated with *agreement type*, and PASS leverages 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 (Bafumi et al. 2005, 176–79).

$$\beta_{type} > 0 \tag{10}$$

However, *agreement type* does not cleanly separate strong and weak agreements.[2] The model thus includes a second identification restriction where the discrimination parameters are constrained to be positive under the assumption that the presence of a provision always signals a stronger agreement.[3] Even if the targeted incompatibility for a provision is not present in a conflict, the inclusion of that provision should fail to increase the strength of an agreement, not decrease it. The error parameter ϵ further addresses this possibility, so this is an innocuous assumption

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

[3]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.

$$\gamma > 0 \tag{11}$$

2.1 Model validation

The identification restriction in Equation 10 needs to be evaluated to assess the validity of the model.[4] If *agreement type* is a good separator of strong and weak agreements, then the estimate for $\beta_{\texttt{type}} \gg 0$. The posterior mean for $\beta_{\texttt{type}}$ is 1 with a 95% credible interval of [0.72, 1.3], so *agreement type* clearly serves to its purpose as an identification restriction.[5] While all other $\boldsymbol{\beta}$ values are positive, only the 95% credible intervals for Protection Measures and Reparations exclude 0.

Figure 2 presents the posterior means and 95% credible intervals for the difficulty parameters $\boldsymbol{\alpha}$ and the discrimination parameters $\boldsymbol{\gamma}$. Higher difficulty parameters indicate provisions that are more effective at differentiating between weak and strong agreements. Difficulty parameters define a 'baseline' of strength that agreements must surpass for there to be a reasonable chance of observing that provision. Ceasefire provisions have an estimated α value of 0.01 and γ value of 1.32 so an agreement does not have to be particularly strong to exhibit a ceasefire provision, while the absence of a ceasefire provision tells us that an agreement is likely to be weak. There is a clear block of provisions including territorial autonomy, local power sharing, and independence referenda with higher difficulty parameters. This high difficulty makes sense given the finding that territorial conflicts are more difficult to peacefully resolve (Toft 2003).

There is also a clear block of provisions such as amnesty for rebels, prisoner release, and peacekeeping with lower difficulty parameters that is agnostic towards the incompatibility in a conflict. As this last group of provisions addresses the consequences of violence rather than underlying incompatibilities responsible for the outbreak of conflict in the first place, they will contribute less to the strength of an agreement. The low α estimates for these parameters reflect exactly this; an agreement does not have to be strong to have a high probability of containing these consequences-of-violence provisions

Just as the identification restriction on $\beta_{\texttt{type}}$ appears justified, constraining $\gamma > 0$ is defensible because

[4]All results are from models with 4 chains run for 20,000 iterations with 15,000 warmup iterations. Inference is performed on the 5000 post-warmup iterations pooled across chains.

[5]Starting values for $\boldsymbol{\theta}$ 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.

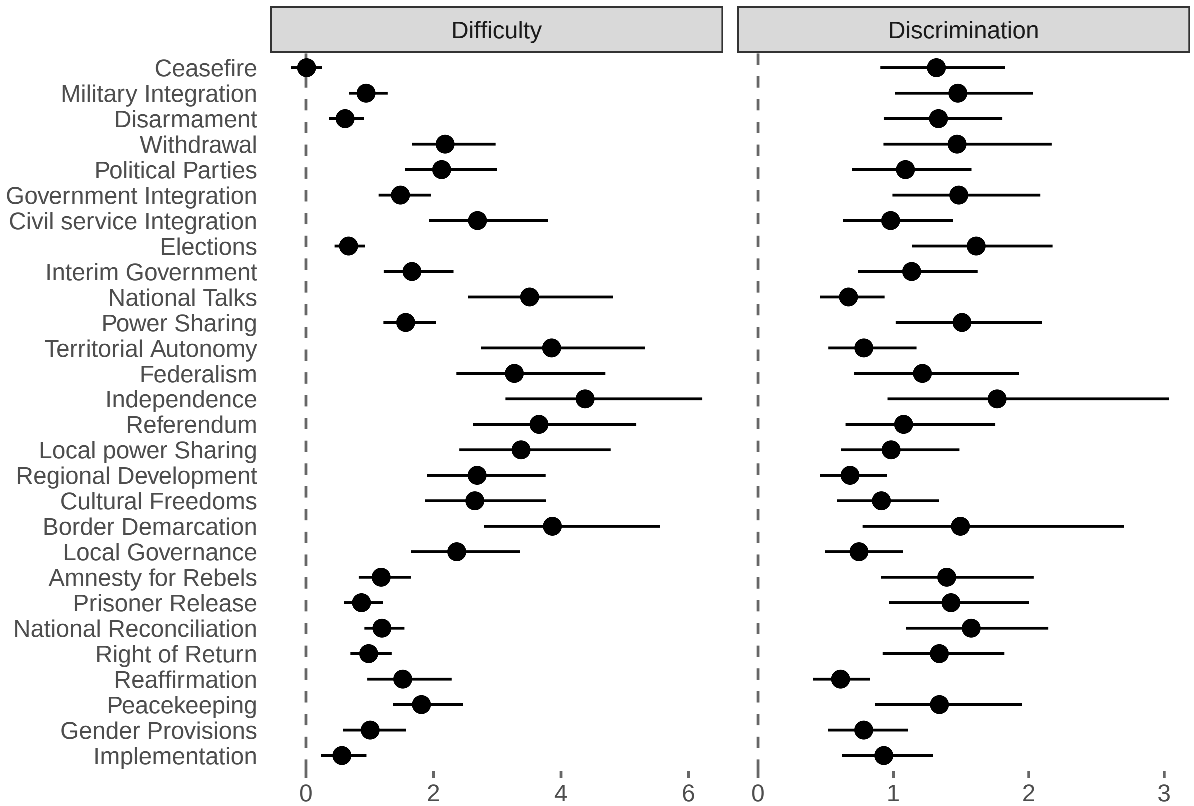

Figure 2: Difficulty and discrimination parameter estimates with 95% credible intervals

none of the credible intervals for γ in Figure 2 approach 0. 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 1 evaluated across the range of $\hat{\boldsymbol{\theta}}$ using $\hat{\alpha}_j$ and $\hat{\gamma}_j$. Figure 3 depicts the ICCs for elections and reaffirmation provisions, along with 95% posterior uncertainty, as well as observed values of y_{ij} for the provisions.[6]

The ICC for elections is much steeper than for reaffirmation because while the observed instances of elections ($y_{ij} = 1$) and no elections ($y_{ij} = 0$) overlap, there is a clear space to the left where the weakest agreements do not contain election provisions with a similar gap to the right where none of the strongest agreements lack election provisions. In contrast, agreements across the range of $\boldsymbol{\theta}$ both have and do not have reaffirmation provisions, so the slope of the ICC is much lower and it does not effectively discriminate between weak and strong agreements.

One way to explicitly model the fact that different provisions are more or less relevant in different conflicts

[6] See the Supplemental Information for similar plots for all 28 provisions.

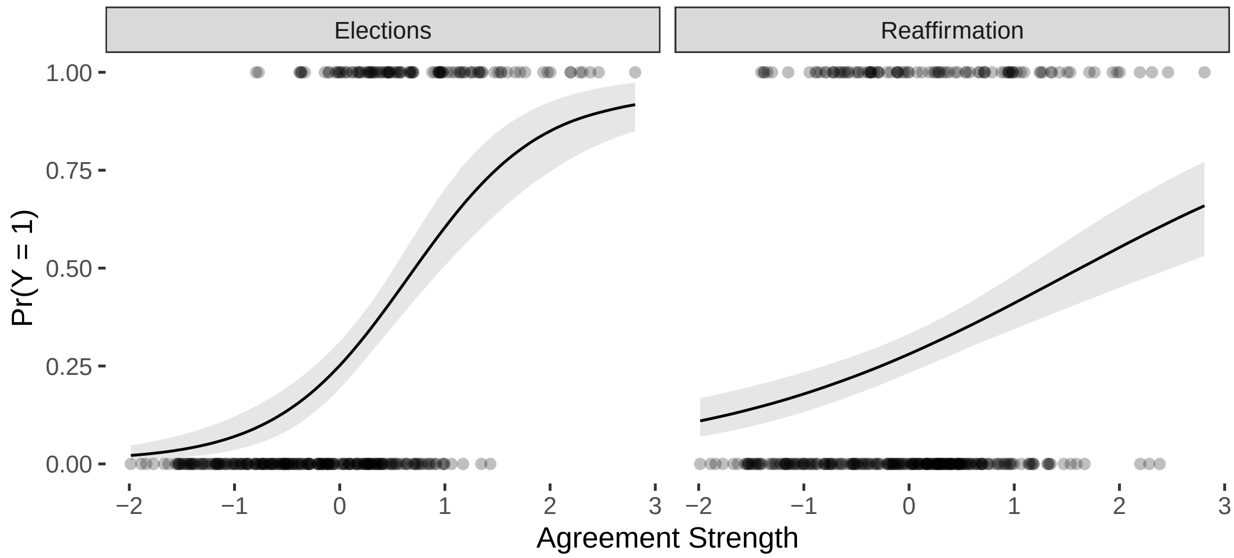

Figure 3: Item characteristic curves and observed provisions with 95% credible intervals

due to varying issue saliences across conflicts would be to allow for differential item functioning in the model. This would let α and γ 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, local power sharing and 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. I instead rely on the conflict random intercept $\boldsymbol{\delta}$ to model some of this heterogeneity.

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 (Toft 2003) due to the strategic or identity value of territory (Toft 2014), so we might expect difficulty parameters to be higher in these conflicts. This model includes two sets of α and γ 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 baseline PASS model classifies 84.42% of observed indicators correctly, while the model with differential item functioning classifies only 81.75% correctly. This decrease in accuracy suggests that the relationship between observed provisions and agreement strength does not significantly vary by incompatibility, and the extra 56 parameters that this model has to estimate reduce its accuracy. Similarly, omitting the PA-X provisions from the prior on $\boldsymbol{\theta}$ decrease classification accuracy to 84.38%.

The error parameter ϵ_0 is estimated to be 0.01 and ϵ_1 is estimated to be 0.05, so the minimum probability of observing a given provision is 0.01 and the maximum is 0.95. Given these relatively low error rates, omitting $\boldsymbol{\epsilon}$ in Equation 1 decreases classification accuracy slightly to 84.35%. While this decrease is minimal, I use the three parameter model to calculate PASS as $\boldsymbol{\epsilon}$ reduces the chances of overfitting to the observed data and increases generalizability.

Finally, we can qualitatively inspect the data to assess the validity of the strength measure. Figure 4 presents provisions in the 10 strongest agreements, in comparison with the selected agreements in Figure 1. 25 agreements have zero provisions, but only 17 of these provision-less agreements appear in the 25 weakest agreements because they are signed in conflicts where other, stronger, agreements are signed, so the conflict random intercept $\boldsymbol{\delta}$ moves their $\boldsymbol{\theta}$ estimates upwards to account for omitted variables likely to be correlated across agreements within the same conflict.

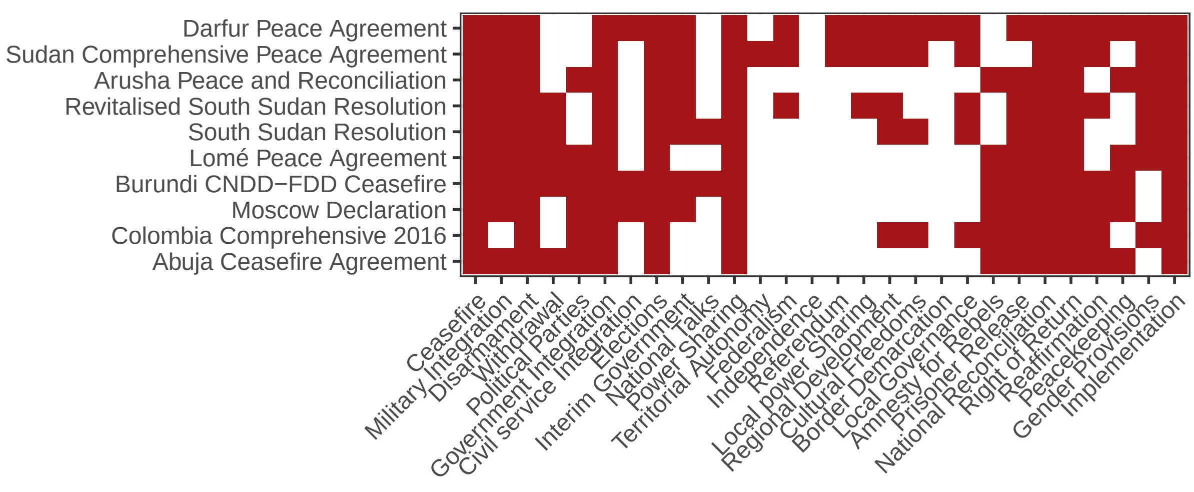

Figure 4: Provisions in ten strongest peace agreements

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 extremely low difficulty and relatively high discrimination of ceasefire provisions 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.

Measuring the predictive accuracy of PASS is another way to validate the scores and illustrate their utility compared to existing measures. I do so for two quantities of interest: whether an agreement ends in failure and how long it survives until such (potential) failure. The former is modeled with a logistic regression while the latter employs a Cox proportional hazards model. Both outcomes use *incompatibility*, *cumulative intensity*, and *cold war* as control variables due to their use in studies of these outcomes. 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 median duration of agreements in the held out set. Both of these metrics assess the predictive accuracy of a model

This process involves generating out of sample predictions to guard against overfitting of the predictive model of the observed data (Hastie, Friedman, and Tibshirani 2009, 219–257;@Ward2010). I carry out this process by conducting 3-fold cross-validation where the data are partitioned into 3 subsets. For each outcome the appropriate model is fit to 2 subsets of the data and then 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 the relevant metric, and the process is repeated nine more times 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.

I compare three alternative measures of agreement strength to PASS. *Power sharing* has been used in multiple studies of peace duration to capture agreement strength (Hartzell, Hoddie, and Rothchild 2001; Hartzell and Hoddie 2003). I also test the predictive power of *agreement type* as comprehensive agreements resolve more incompatibilities than partial agreements, and both resolve more than process ones. Finally, Werner and Yuen (2005) create an *additive index* of agreement provisions to measure strength, so I evaluate one as well. I perform 3-fold cross-validation to assess the predictive accuracy for all four measures of strength for both agreement outcome and duration. Table 4 presents these results for both tasks.

Predictor	AUC	IBS
Power Sharing	0.5278	0.6994
Agreement Type	0.5287	0.7894
Additive Index	0.5526	0.6969
PASS	0.5583	0.6955

Table 4: Average area under the curve (higher is better) and integrated Brier Score (lower is better) for 3-fold cross-validation

PASS performs best at both the binary and duration prediction tasks. In each case, the *additive index* performs better than *agreement type*, while PASS does better still. This pattern suggests that PASS offers improvements over existing measures of peace agreement strength as a predictor in substantive analyses, and that increasing methodological complexity yields measures that better approximate the strength of each agreement.

3 Peace agreement strength scores

With the model validated, I now present the distribution of PASS. Figure 5 displays the posterior mean of each agreement's strength, as well as its 95% credible interval with the agreements described in Figure 1 labeled.

Some initial observations provide face validity for the estimates. 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, 0 agreements have identical $\boldsymbol{\theta}$ values due to the information sharing accomplished by $\boldsymbol{\delta}$.

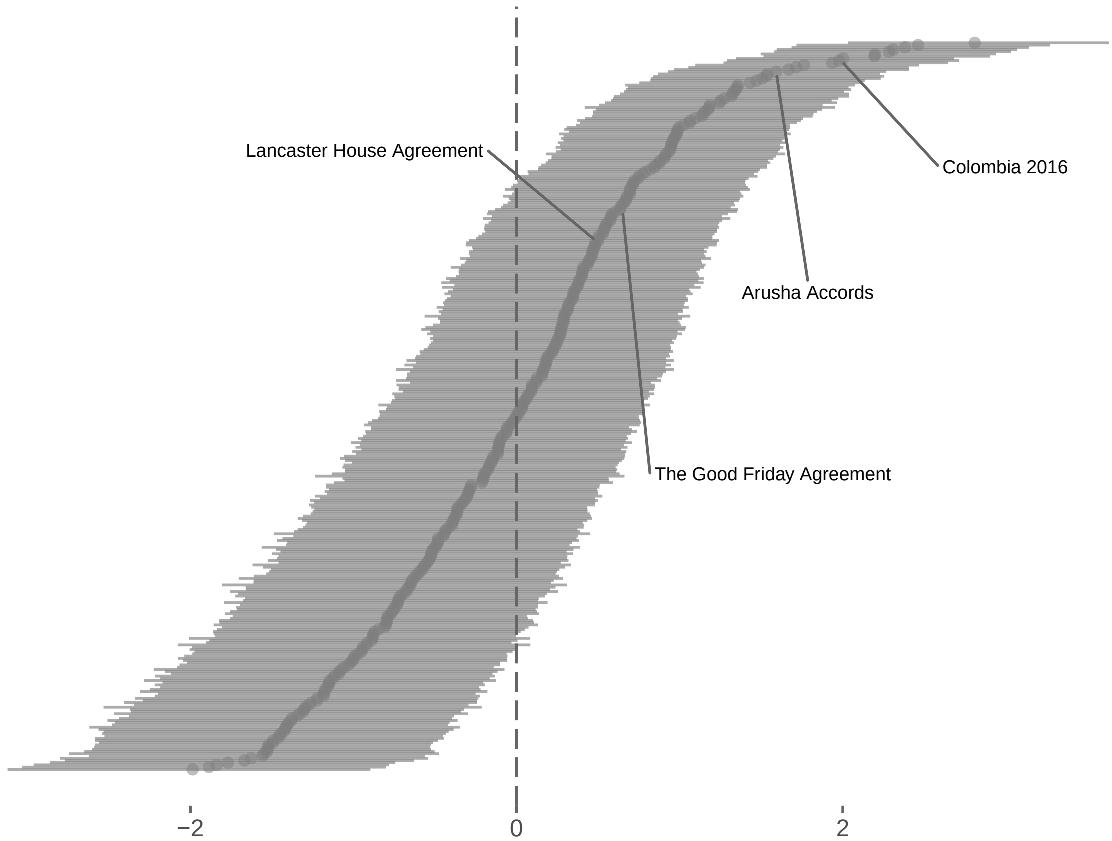

Figure 5: Distribution of agreement strengths with 95% credible intervals

4 Selection effects in agreement signing

A direct measure of peace agreement strength allows us to investigate whether selection effects that apply to certain specific provisions within peace agreement strength apply to agreement strength more broadly. 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.

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 PASS. The higher the death toll in a conflict, the longer and fiercer it has been (Balch-Lindsay and Enterline 2000), which I capture with (the natural log of) *cumulative battle deaths* from the year of signing from the UCDP Battle-Related Deaths Dataset (Pettersson, Högbladh, and Öberg 2019). I directly measure the length of a

	Model 1	Model 2	Model 3	Model 4
ln(Cumulative Battle Deaths)	0.07			
	(0.03)			
Active Years		-0.17^{*}		
		(0.04)		
Internationalized			0.47^{*}	
			(0.14)	
Ethnic Conflict				0.65^{*}
				(0.13)
(Intercept)	-0.40	0.40^{*}	-0.07	-0.55^{*}
	(0.29)	(0.11)	(0.05)	(0.12)
R^2	0.02	0.04	0.04	0.07
Adj. R^2	0.01	0.04	0.03	0.06
Num. obs.	211	328	328	328
RMSE	0.90	0.88	0.89	0.87

$^{*}p < 0.05$

Table 5: Linear models of conflict severity and agreement strength

conflict by counting the number of *active years* before an agreement is signed. Conflicts with many combatants take longer to resolve due to the increased number of veto players (Cunningham 2006), so civil wars that have been *internationalized* may also have stronger agreements due to the need to satisfy more stakeholders. 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). Table 5 presents results from these models.

The relationship between *cumulative battle deaths* and agreement strength is positive but not statistically significant. However, the battle deaths data do not begin until 1989 and are missing for 32.37% of observations 1989 or later, so this result should be viewed with skepticism. Longer conflicts appear to be associated with weaker agreements as *active years* is negative and significant. The *internationalized* and *ethnic conflict* variables are positive and significant, suggesting that stronger agreements are more likely in these less easily resolved conflicts. Taken together these findings suggest that the weaker agreements may be signed in longer conflicts, but that harder to resolve conflicts are associated with stronger agreements.

5 Conclusion

PASS comes ready to employ in analyses of conflict and conflict resolution. The data contain conflict, year, and dyad identifiers compatible with the UCDP/PRIO ACD version 19.2 data. These scores allow researchers to 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.

The utility of this new measure is illustrated by the ability to identify a selection effect in the design of peace agreements; stronger agreements are more likely to be signed in harder to resolve conflicts. 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.

6 References

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.

Balch-Lindsay, Dylan, and Andrew J. Enterline. 2000. "Killing Time: The World Politics of Civil War Duration, 1820-1992." *International Studies Quarterly* 44 (4): 615–42.

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

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.

Carter, Jeff, and Charles E. Smith Jr. 2019. "A Framework for Measuring Leaders' Willingness to Use Force." Working Paper.

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.

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.

Fortna, Virginia. 2008. *Does Peacekeeping Work? Shaping Belligerents’ Choices After Civil War*. Princeton: 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.

Harbom, Lotta, Stina Högbladh, and Peter Wallensteen. 2006. “Armed Conflict and Peace Agreements.” *Journal of Peace Research* 43 (5): 617–31.

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

Hartzell, Caroline A., and Matthew Hoddie. 2015. “The Art of the Possible: Power Sharing and PostCivil War Democracy.” *World Politics* 67 (1): 37–71.

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.

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.

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.

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

Licklider, Roy. 1995. “The Consequences of Negotiated Settlements in Civil Wars, 1945-1993.” *The American Political Science Review* 89 (3): 681–90.

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

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.

Nilsson, Desirée. 2008. “Partial Peace: Rebel Groups Inside and Outside of Civil War Settlements.” *Journal of Peace Research* 45 (4): 479–95.

Pettersson, Therése, Stina Högbladh, and Magnus Öberg. 2019. “Organized Violence, 19892018 and Peace Agreements.” *Journal of Peace Research*, June, 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.

Svensson, Isak. 2007. “Mediation with Muscles or Minds? Exploring Power Mediators and Pure Mediators in Civil Wars.” *International Negotiation* 12 (2): 229–48.

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.

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.

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

Walter, Barbara F. 2009. “Bargaining Failures and Civil War.” *Annual Review of Political Science* 12 (1): 243–61.

———. 1997. “The Critical Barrier to Civil War Settlement.” *International Organization* 51 (03): 335–64.

Werner, Suzanne, and Amy Yuen. 2005. “Making and Keeping Peace.” *International Organization* 59 (2): 261–92.

Williams, Rob, Daniel J. Gustafson, Stephen E. Gent, and Mark J. C. Crescenzi. n.d. “A Latent Variable Approach to Measuring and Explaining Peace Agreement Strength.” *Political Science Research and Methods*, 1–17.