Supplemental Information for Peace Agreement Strength Scores

Rob Williams*

December 16, 2020

Contents

*Postdoctoral Research Fellow, Washington University in St. Louis, rob.williams@wustl.edu, jayrobwilliams.com

1 Descriptive statistics

Figure 1 displays the frequency of all agreement provisions **Y**.

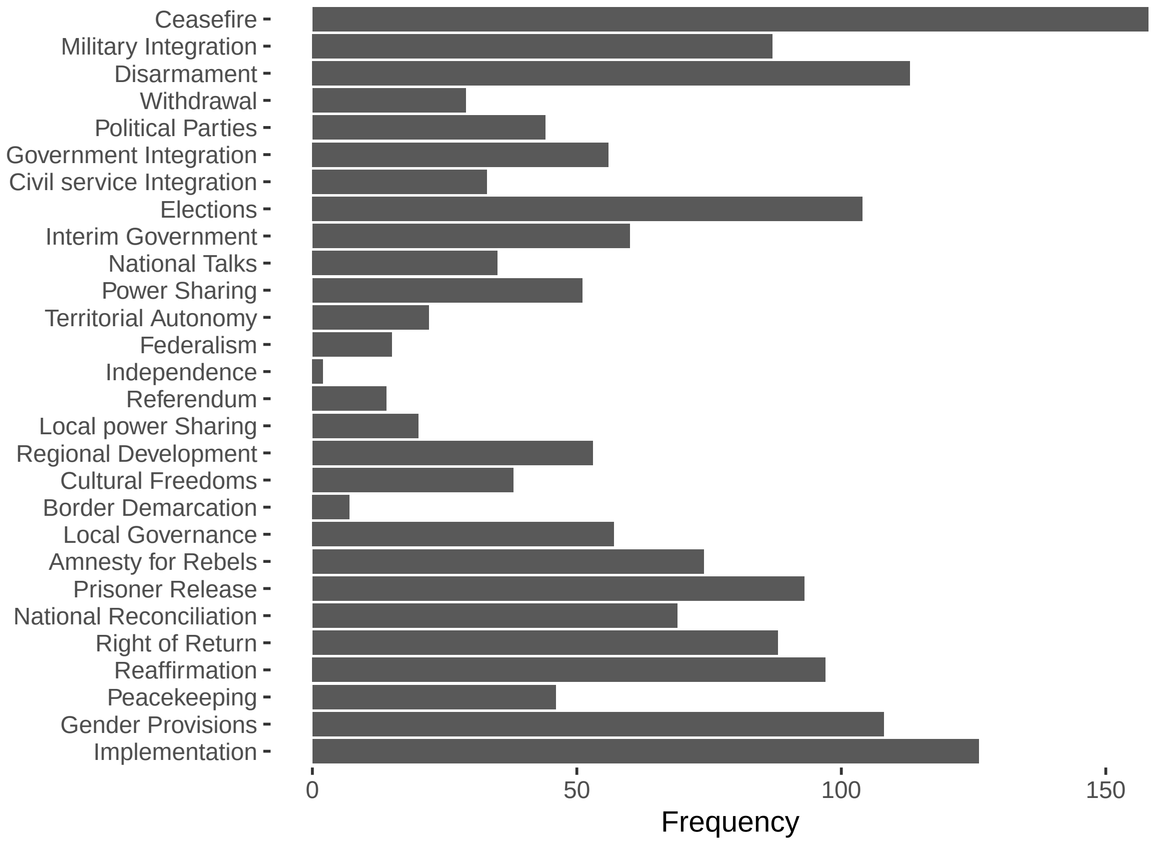

Figure 1: Count of provisions in the data.

Figure 2 displays the correlation matrix for agreement provisions **Y**.

Figure 3 displays the correlation matrix for agreement-level covariates including *agreement type* (Harbom, Högbladh, and Wallensteen 2006; Pettersson, Högbladh, and Öberg 2019) and the PA-X provisions (Bell and Badanjak 2019) contained in **X**.

Figure 4 illustrates the data coverage across the three data sources used in PASS.

1.1 Multi-conflict agreements

An agreement can be signed to terminate multiple separate conflicts, and the UCDP Peace Agreements Data contain 3 such agreements that are signed in more than one conflict. Table 1 presents these agreements. I deal with these cases by splitting the agreements, creating one observation per agreement-conflict pair (e.g.,

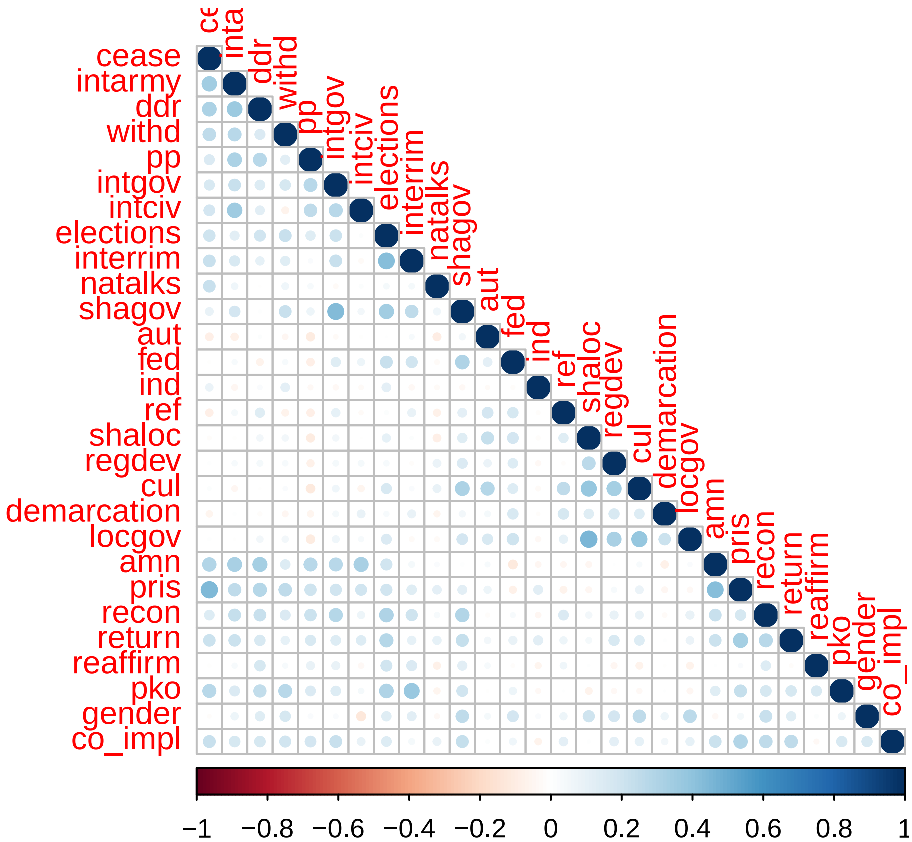

Figure 2: Correlation matrix for agreement provisions

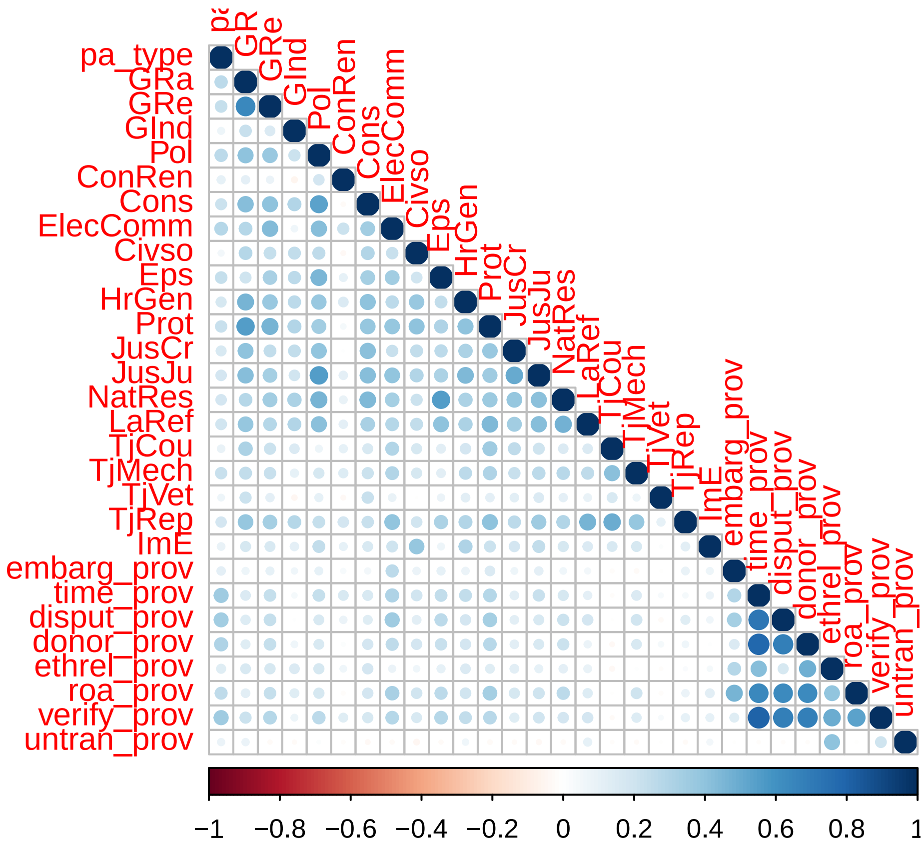

Figure 3: Correlation matrix for covariates on θ prior components

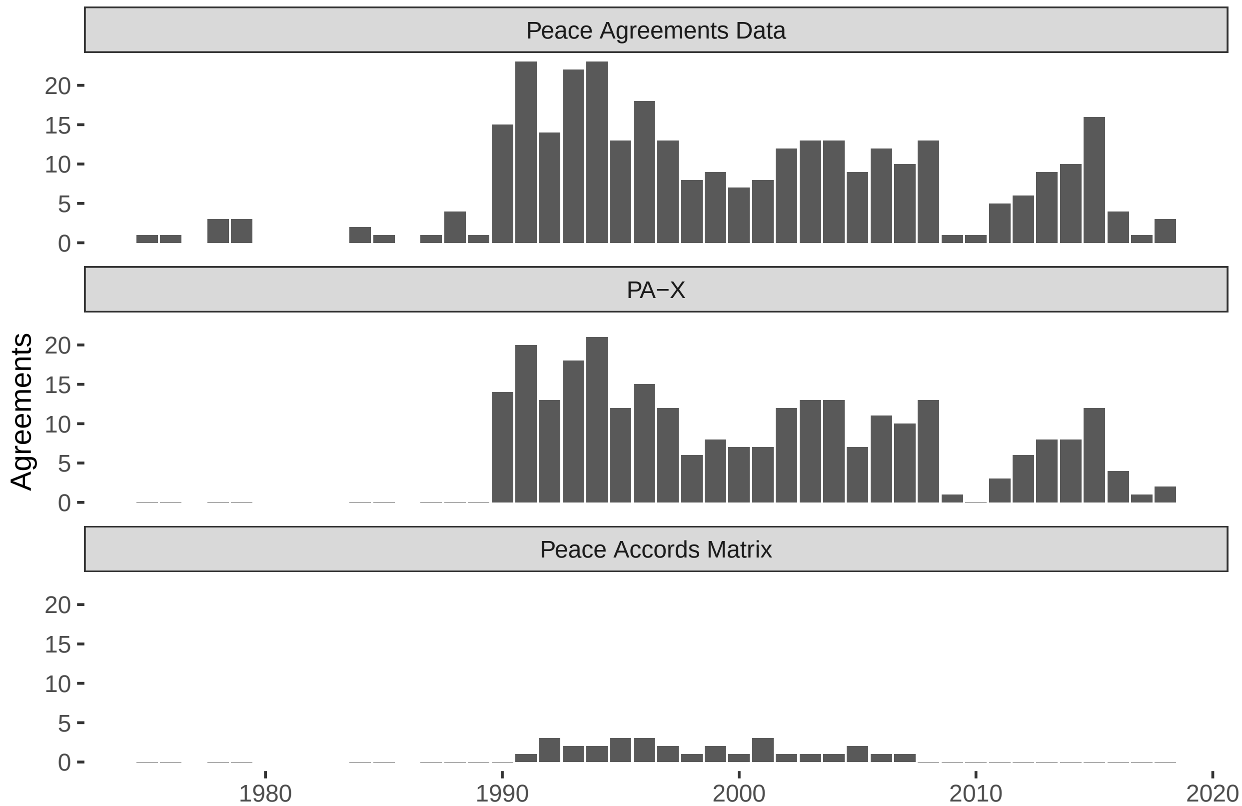

Figure 4: Coverage across data sources

the Vance-Owen Plan in the Bosnia-Herzegovina and Serb conflict is separated from the Vance-Owen Plan in the Bosnia-Herzegovina and Croat conflict). The Vance-Owen Plan and and Deed of Commitment thus become two separate agreements, while the Nationwide Ceasefire in Myanmar becomes three. This splitting is necessary because the same agreement may be stronger or weaker in different conflicts due to different underlying issues driving the violence or different drivers of post-conflict instability.

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 1: Multiple conflict agreements

An agreement with the same conflict resolution 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. Similarly, the same conflict prevention provision may be varyingly effective in different dyads involved in the same conflict. A group that has an external ally that can deter the government from reneging on an agreement will benefit less from detailed enforcement mechanisms than a group without such an ally, meaning that the contribution of detailed enforcement mechanisms to agreement strength will be lower in the former case.

Splitting agreement signed in multiple conflicts also makes empirical 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).

Splitting the multi-conflict agreements introduces 3 sets of agreements with identical provisions. The model will give each set of disaggregated agreements identical θ values as the data used to estimate them will be identical. While it may seem problematic that agreements will have identical strength estimates even though they address different contexts, this is actually desirable. Because PASS uses only the content of agreements themselves, an agreement signed to terminate two different conflicts will have the same strength in both conflicts. To assess the independent effect of peace agreements on post-conflict outcomes, we must account for all other relevant factors, but doing so requires a measure of agreement strength that does not draw on outside information.

2 Model parameters

Figure 5 presents the item characteristic curves and observed values for all provisions in PASS, replicating Figure 3 for all provisions.

3 Cross-validation

The 3-fold cross-validation uses fits two different types of models to outcome data. For agreement outcome (continuing or failed) it uses logistic regression and for agreement duration it uses Cox proportional hazard regression. In both cases, it includes a dummy variable to account for whether an agreement was signed

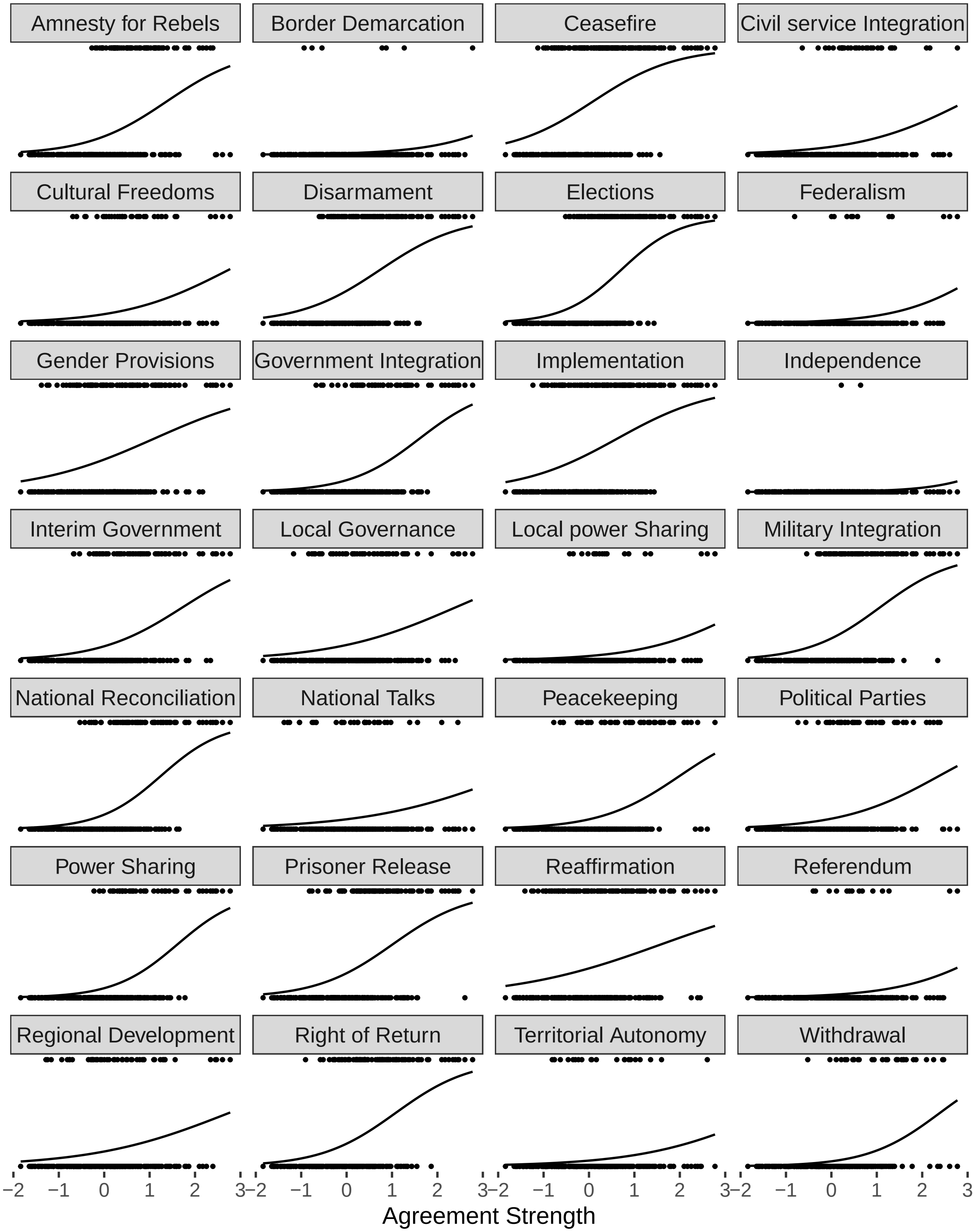

Figure 5: Distribution of observed provisions and item characteristic curves

during the *Cold War* or not.

4 Sensitivity analyses

4.1 Provision selection

The identification restriction that $\beta_{type} > 0$ requires evaluating whether any provisions should be excluded due to being related to a different latent construct. This is done by examining whether any γ_j estimates have posterior distributions close to 0 (Bafumi et al. 2005). Figure 6 displays the posterior distributions for all $\boldsymbol{\gamma}$ estimates.

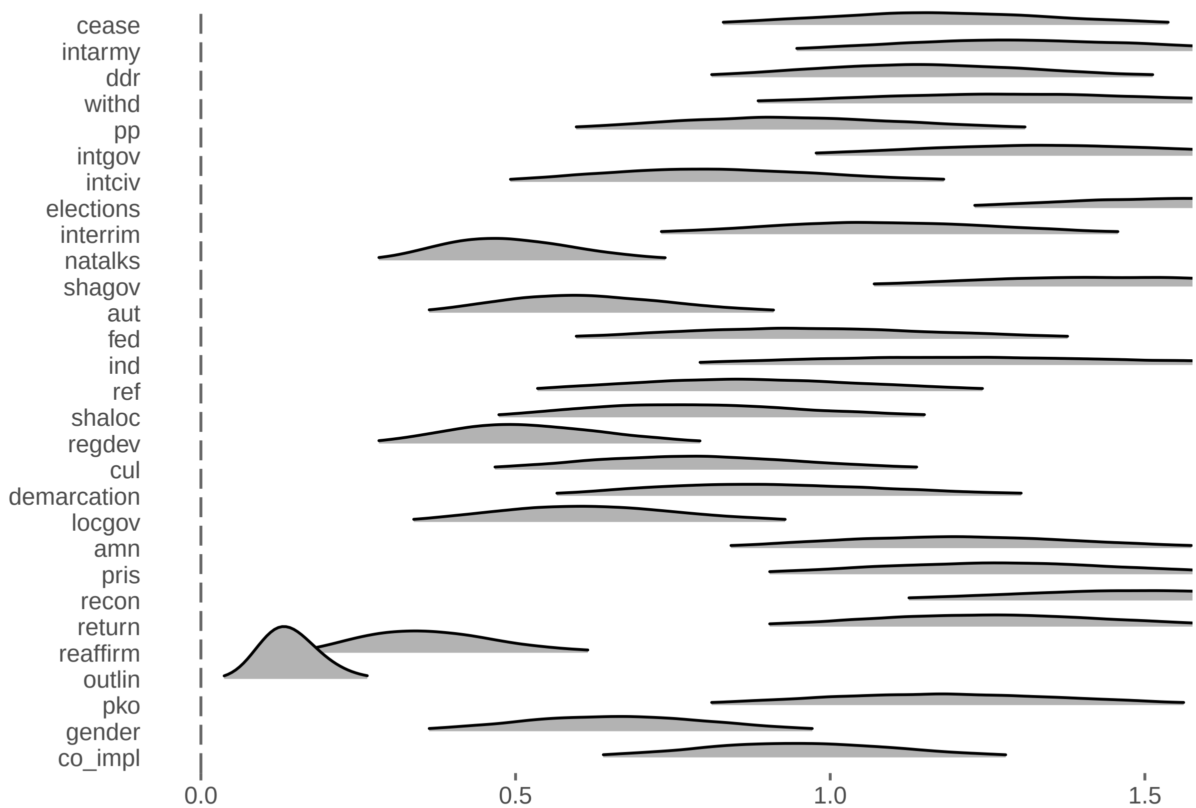

Figure 6: Posterior densities for all discrimination parameters

The only provision with a density closer to 0 is outlining. The group of provisions that had posterior densities close to 0 and were ommitted from Williams et al. (n.d.) (autonomy, federalism, independence, referendum, local power sharing, regional development, cultural freedoms, and local governance) are discernable as a group of provisions with a lower average $\boldsymbol{\gamma}$ value than the retained provisions, with the exception of independence which is much higher than in Williams et al. (n.d.). However, none of these provisions have distributions suggesting their exclusion. When including the outlining provision, only 6 agreements have no provisions. This is in contrast to 25 agreements with no provisions in the paper when outlining is excluded.

Figure 7 plots the rank ordering of agreement strength for the main PASS model presented in the paper as well as one that includes the outlining provision.

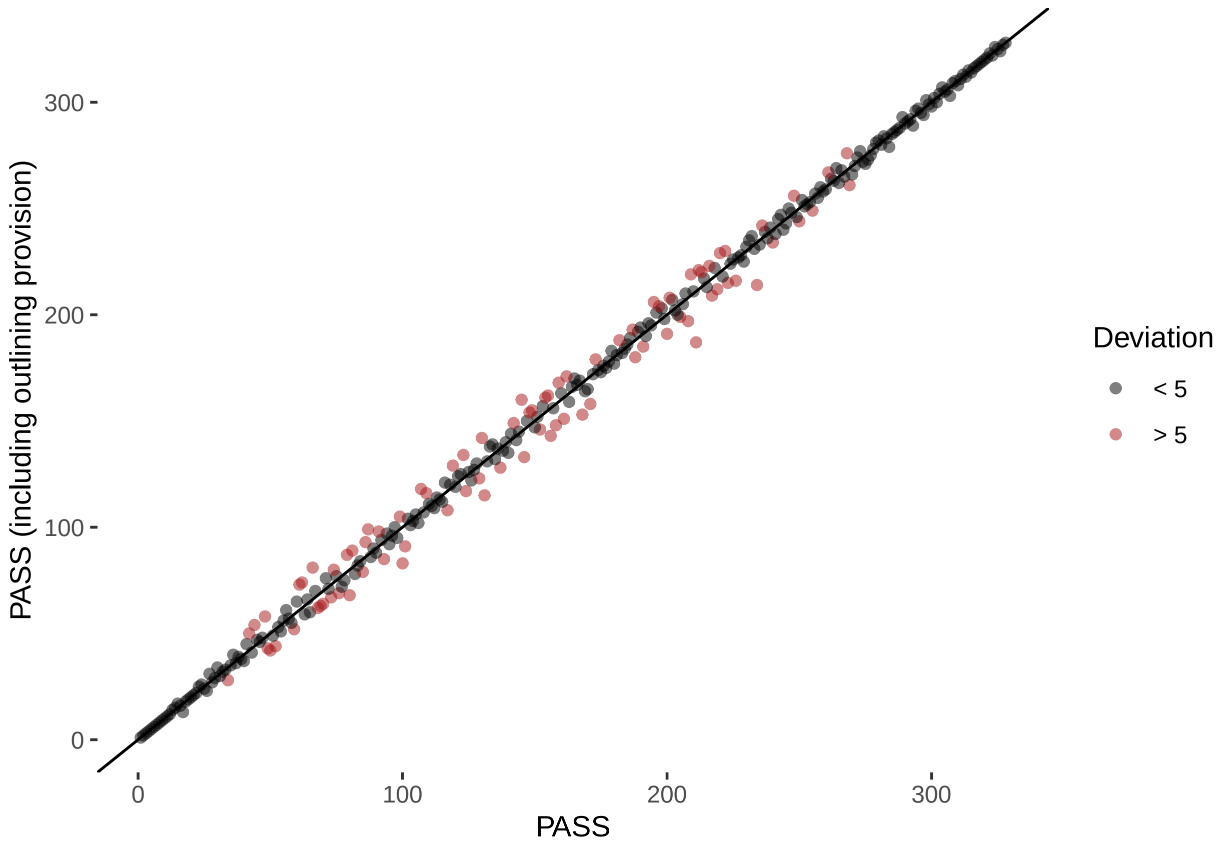

Figure 7: Shift in rank ordering of agreement strengths between PASS and a model using all provisions

Figure 8 replicates Figure 6 in the paper but includes the outlining provision.

4.2 Robust IRT model

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. To account for this data sparsity, I

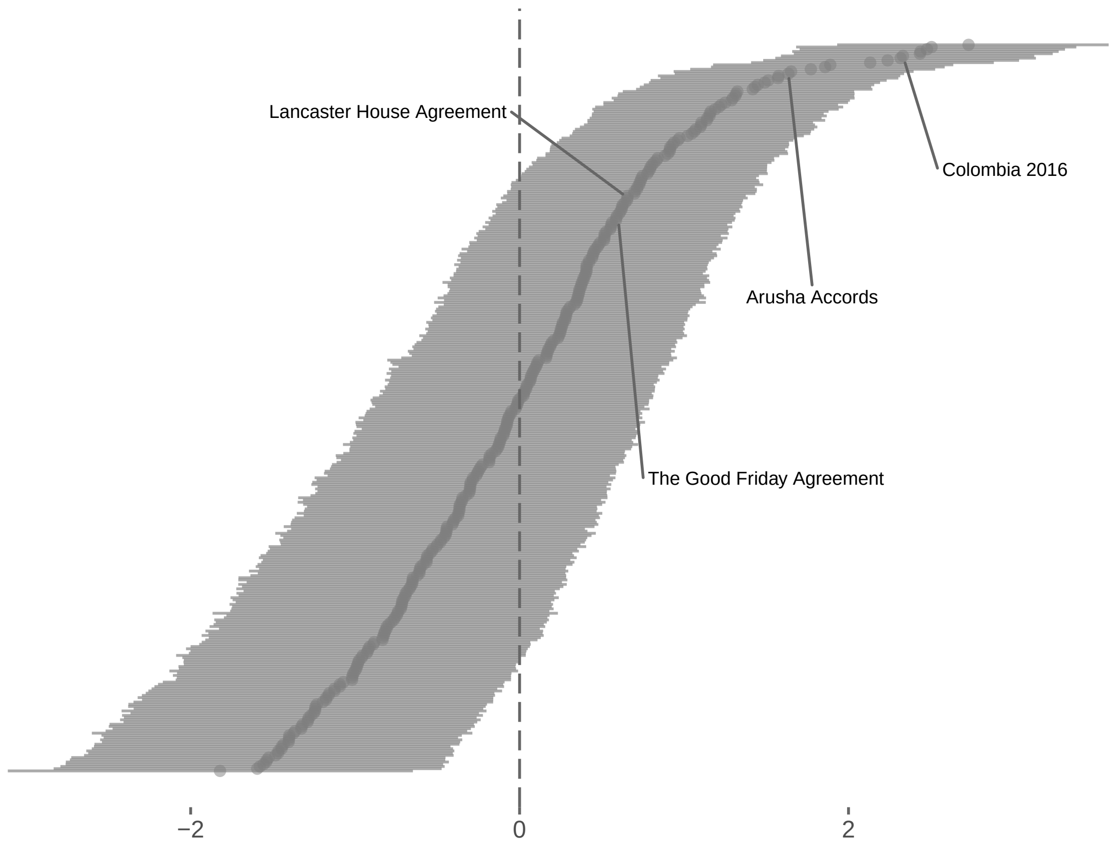

Figure 8: Distribution of agreement strengths with all provisions

consider an extension of the two parameter item response model that introduces a third 'error' parameter, $\boldsymbol{\epsilon}$,

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

which accounts for this lack of separability. 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 0.06, so the minimum probability of observing a given provision is 0.01 and the maximum is 0.94. Given these relatively low error rates, the three parameter IRT model in Equation 1 has a classification accuracy (84.36%) indistinguishable from the two parameter model (84.4%). Given the equal predictive accuracy, I use the more parsimonious two parameter model.

4.3 Omitting PA-X and PAM covariates

Figure 9 plots the rank ordering of agreements in both scores aganist one another. Any agreement whose rank order position shifts more than five places between the two scores is plotted in red.

Omitting the PA-X covariates yields a classification accuracy of 84.29%.

4.4 Differential item functioning

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, local power

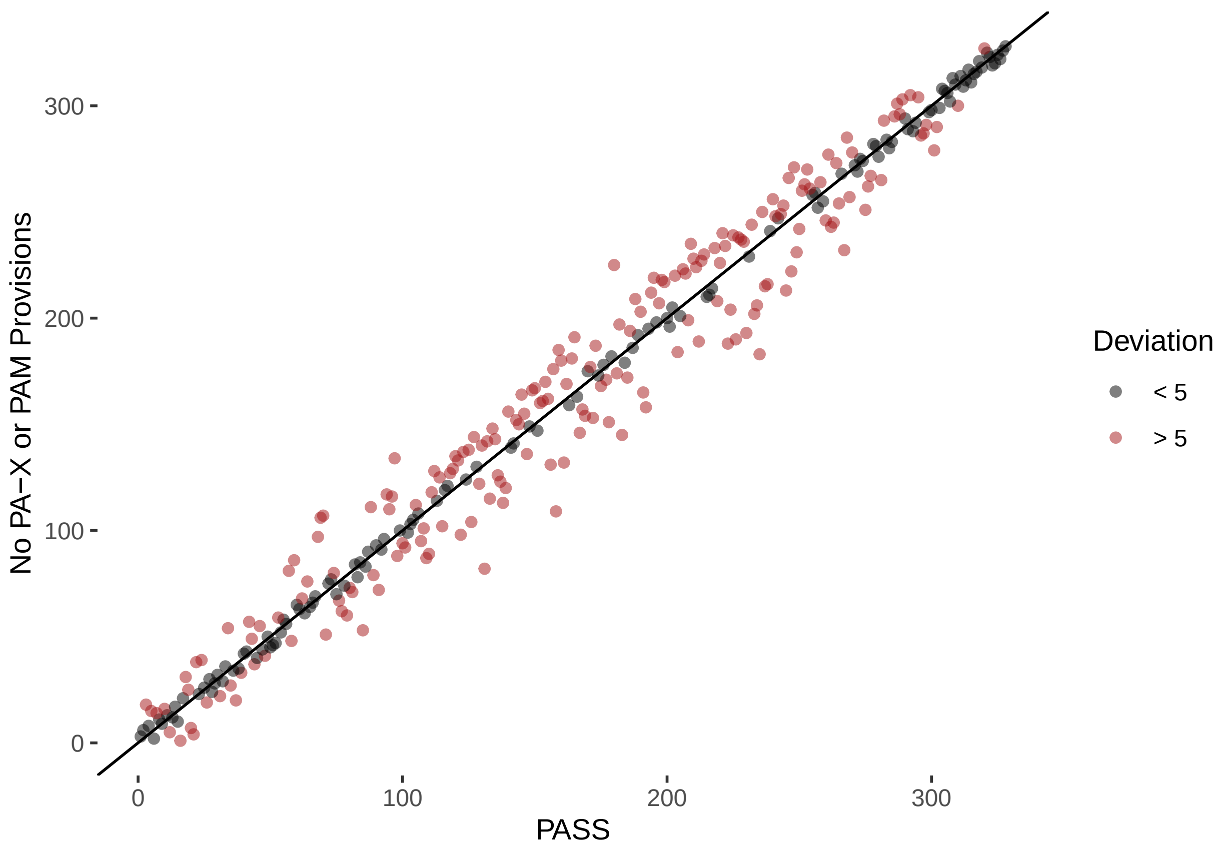

Figure 9: Rank odering of agreement strengths between the full scores and those with no information from PA-X or PAM

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.

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 $\boldsymbol{\alpha}$ 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 baseline PASS model classifies 84.4% of observed indicators correctly, while the model with differential item functioning classifies only 76.15% 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. The larger concern is that this strategy relies on information outside that contained in agreements themselves; the type of conflict an agreement was signed in. If we wish to measure the independent effect of peace agreements, then we cannot use conflict-level information of measure the strength of agreements.

Figure 10 illustrates that many provisions such as peacekeeping, local powersharing, and federalism have ICCs that are indistinguishable across conflict incompatibilities. In contrast, civil service integration has much more discriminatory power in governmental conflicts than territorial ones.

4.5 Conflict-level information

An alternative modeling strategy is to add conflict level information to $\mathbf{X}$'s contributions to the prior on θ: *incompatibility*, whether the agreement was signed in an *active conflict* year, and number of *previous agreements* in the conflict. Including this conflict-level information reduces classification accuracy to 84.34%. It also

Figure 10: Distribution of observed provisions and item characteristic curves with differential item functioning by conflict incompatibility

renders the scores unusable for any analyses that also include these covariates. By only including information contained within the agreements themselves, PASS can be used in a wide variety of analyses.

5 Advantages over additive index

To illustrate the prevalence of ties that an additive index would yield, Figure 11 presents a histogram of additive index values for all 328 agreements in the data.

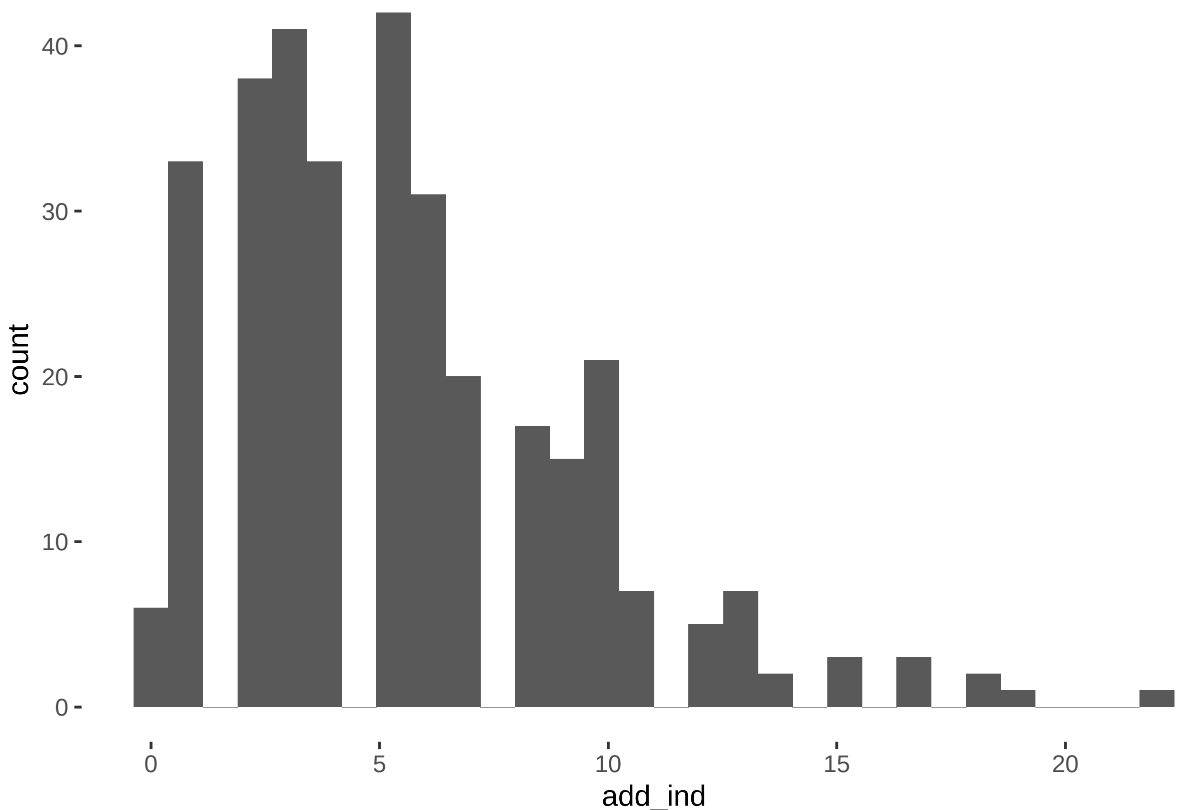

Figure 11: Histogram of additive index values

Although Williams et al. (n.d.) find that their latent measure of agreement strength is highly correlated with a simple additive index of provisions, PASS has many advantages over an additive index. With these updated data, no agreement has more than 22 of 29 provisions. However, many agreements have the same number of provisions, so a latent variable approach to measuring agreement strength solves the problem of ties in the additive index. The most common number of provisions, 5, occurs in 42 agreements.

Any analysis that explains changes in the strengths of peace agreements over the duration of a conflict can incorporate uncertainty about agreement strength in a way that an additive index cannot.

6 Duration

The Cox proprotional hazard models mentioned in the conclusion are presented in Table 2. The first two columns use time-invariant covariates, while the third includes the time-varying covariate of aggregate implmentation. All three fail to find a significant relationship between agreement strength and duration.

	Full Sample	PAM Only	PAM Only
Agreement Strength	-0.31^{*}	0.22	0.48
	(0.10)	(0.47)	(0.45)
Aggregate Implementation			-0.02
			(0.02)
AIC	1662.76	66.42	55.27
Num. events	152	10	10
Num. obs.	328	30	272
PH test	0.86	0.90	0.12

$^{*}p < 0.05$

Table 2: Cox proportional hazards models of agreement failure

7 Model estimation with Stan

The IRT parameters θ, α, and γ are reparameterized after estimation in terms of the mean and standard deviation of θ following (Bafumi et al. 2005) to reduce correlation among the IRT parameters and speed up sampling.

$$\theta_i^{adj} = \frac{(\theta_i - \bar{\theta})}{sd(\theta)} \tag{2}$$

$$\alpha_j^{adj} = \frac{(\alpha_j - \bar{\theta})}{sd(\theta)} \tag{3}$$

$$\gamma_j^{adj} = \gamma_j \, sd(\theta) \tag{4}$$

The parameters δ, α, and γ are further reparameterized during estimation with a non-centered parameterization to speed up sampling:

```
data {
  int<lower=1> C;
}
parameters {
  vector[0] alpha_raw;
  vector<lower=.001>[0] gamma_raw;
```

```
  vector[M] theta_raw
  vector[C] delta_raw;
}
transformed parameters {
  vector[O] alpha_reparam;
  vector<lower=.001>[O] gamma_reparam;
  vector[C] delta;
  alpha_reparam = mu_alpha + sigma_alpha * alpha_raw;
  gamma_reparam = mu_gamma + sigma_gamma * gamma_raw;
  delta = mu_delta + sigma_delta * delta_raw;
}
model {
  alpha_raw ~ std_normal();
  gamma_raw ~ std_normal();
  delta_raw ~ std_normal();
}
```

8 MCMC diagnostics

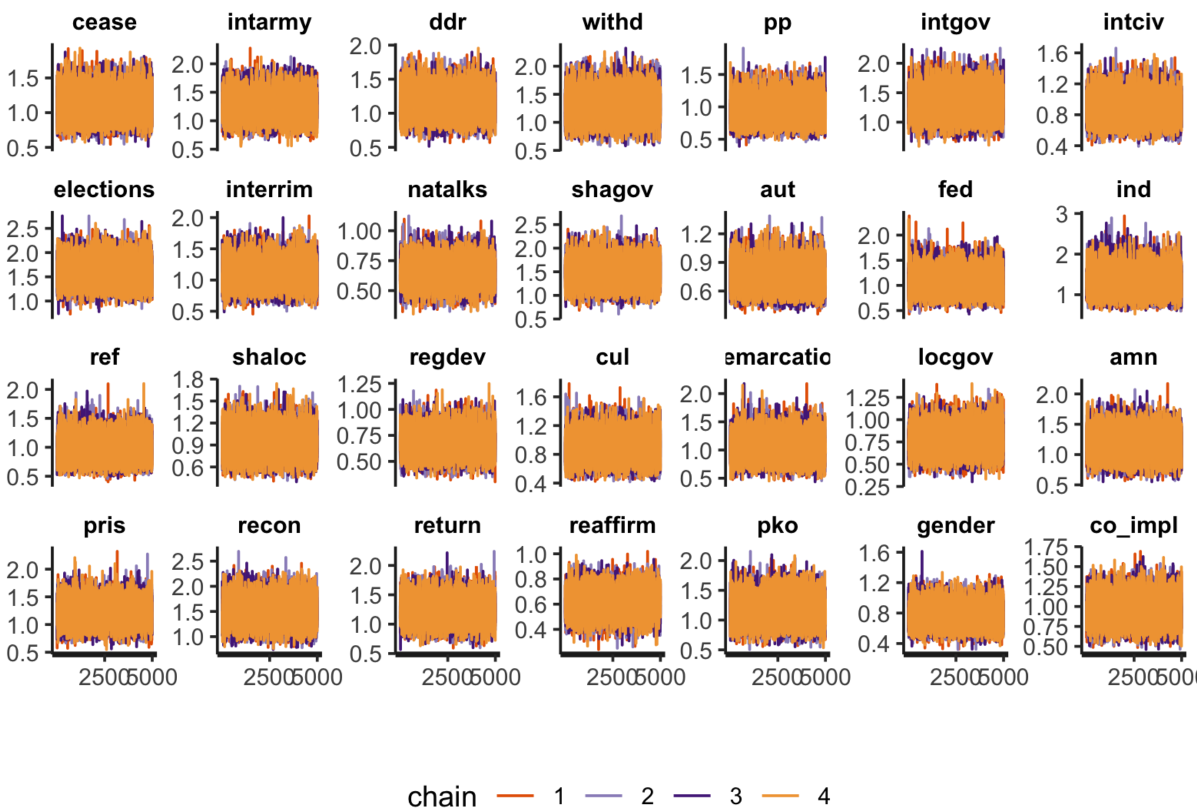

Figure 12: Discrimination parameters

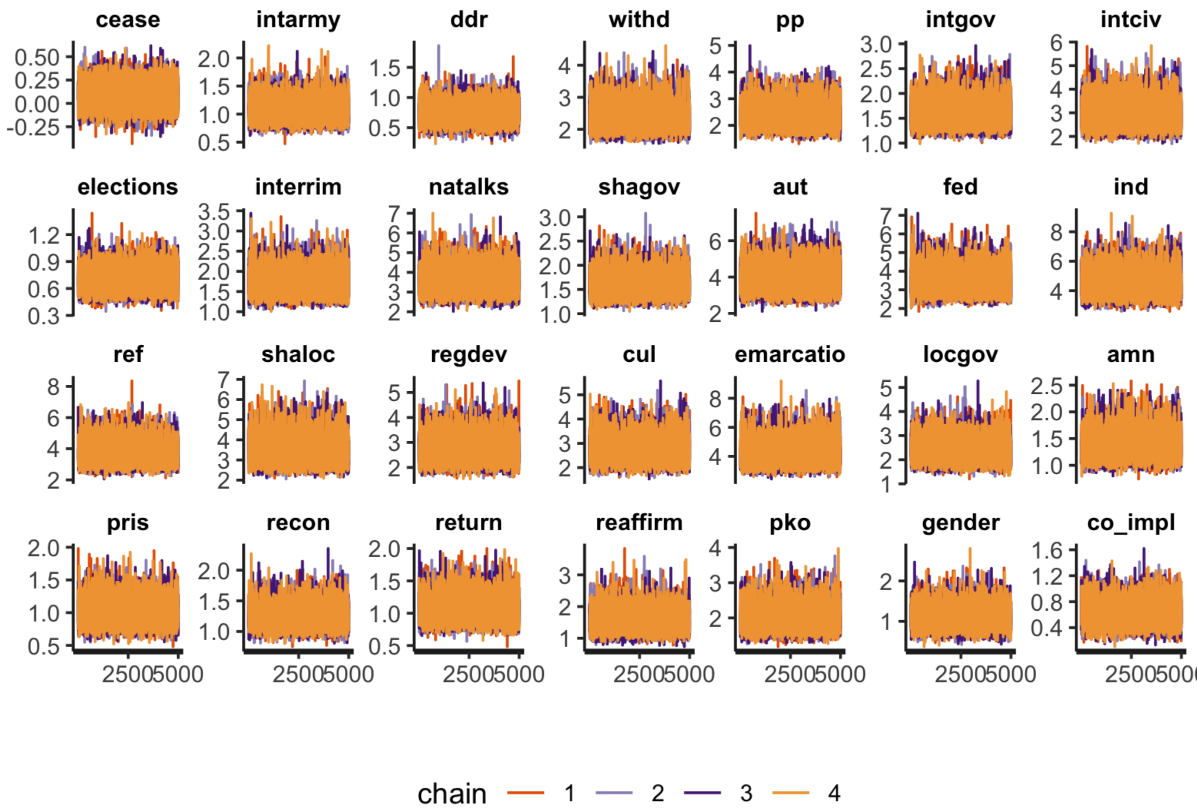

Figure 13: Difficulty Parameters

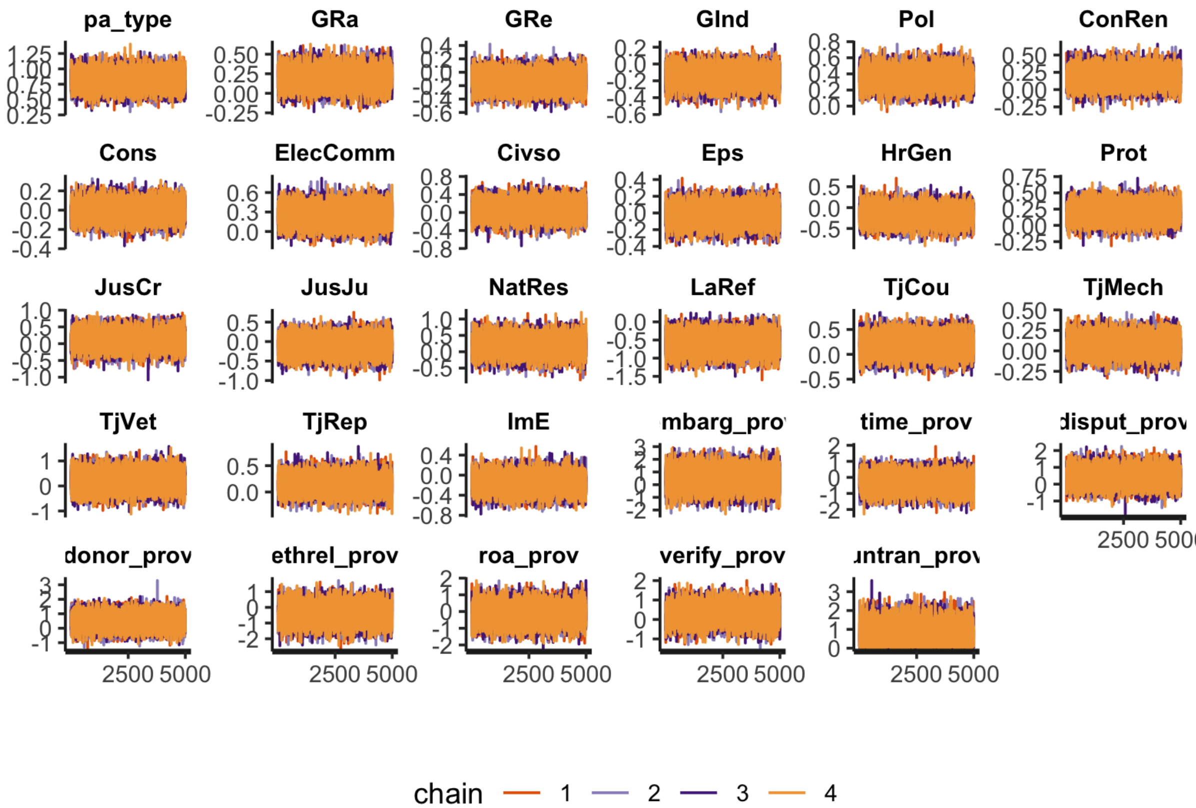

Figure 14: Theta parameters

9 Computing environment

- R version 4.0.2 (2020-06-22), `x86_64-apple-darwin17.7.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Running under: `macOS High Sierra 10.13.6`
- Matrix products: default
- BLAS:

  `/System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/l`
- LAPACK:

  `/System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/l`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: corrplot 0.84, dplyr 1.0.2, english 1.2-5, forcats 0.5.0, ggplot2 3.3.2, ggrepel 0.8.2, ggridges 0.5.2, purrr 0.3.4, readr 1.3.1, rstan 2.21.3, StanHeaders 2.21.0-6, stringr 1.4.0, survival 3.2-3, texreg 1.37.5, tibble 3.0.4, tidyr 1.1.2, tidyverse 1.3.0, xtable 1.8-4
- Loaded via a namespace (and not attached): abind 1.4-5, arm 1.11-2, assertthat 0.2.1, backports 1.1.10, base64enc 0.1-3, blob 1.2.1, boot 1.3-25, broom 0.7.0, callr 3.5.1, cellranger 1.1.0, checkmate 2.0.0, cli 2.1.0, cluster 2.1.0, coda 0.19-4, codetools 0.2-16, colorspace 1.4-1, compiler 4.0.2, crayon 1.3.4, curl 4.3, data.table 1.13.0, DBI 1.1.0, dbplyr 1.4.4, digest 0.6.27, ellipsis 0.3.1, evaluate 0.14, fansi 0.4.1, farver 2.0.3, foreign 0.8-80, Formula 1.2-3, fs 1.5.0, generics 0.0.2, GGally 2.0.0, ggmcmc 1.5.0, ggstance 0.3.4, glue 1.4.2, grid 4.0.2, gridExtra 2.3, gtable 0.3.0, haven 2.3.1, Hmisc 4.4-1, hms 0.5.3, htmlTable 2.0.1, htmltools 0.5.0.9001, htmlwidgets 1.5.1, httr 1.4.2, inline 0.3.16, jpeg 0.1-8.1, jsonlite 1.7.1, knitr 1.30, labeling 0.4.2, lattice 0.20-41, latticeExtra 0.6-29, lifecycle 0.2.0, lme4 1.1-23, loo 2.3.1, lubridate 1.7.9, magrittr 1.5, MASS 7.3-52, Matrix 1.2-18, matrixStats 0.57.0, minqa 1.2.4, modelr 0.1.8, munsell 0.5.0, nlme 3.1-149, nloptr 1.2.2.2, nnet 7.3-14, openxlsx 4.1.5, parallel 4.0.2, pillar 1.4.6, pkgbuild 1.1.0, pkgconfig 2.0.3, plyr 1.8.6, png 0.1-7, prettyunits 1.1.1, processx 3.4.4,

ps 1.4.0, R6 2.4.1, RColorBrewer 1.1-2, Rcpp 1.0.5, RcppParallel 5.0.2, readxl 1.3.1, reprex 0.3.0, reshape 0.8.8, rio 0.5.16, rlang 0.4.8, rmarkdown 2.3, rpart 4.1-15, rstudioapi 0.11, rvest 0.3.6, scales 1.1.1, splines 4.0.2, statmod 1.4.34, stats4 4.0.2, stringi 1.5.3, tidyselect 1.1.0, tools 4.0.2, V8 3.4.0, vctrs 0.3.4, withr 2.3.0, xfun 0.18, xml2 1.3.2, yaml 2.2.1, zip 2.1.1

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.

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

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

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

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, 1989-2018 and Peace Agreements." *Journal of Peace Research*, June, 1–15.

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.

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.