Supplemental Information for A Latent Variable Approach to Measuring and Explaining Peace Agreement Strength

Rob Williams Daniel J. Gustafsone Stephen E. Gent
Mark J.C. Crescenzi

July 6, 2019

Contents

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

Table A.1: Peace agreement provisions in the UCDP peace agreements data, with citations for provisions that are associated with increased agreement duration. We omit border demarcation provisions from our analysis because no agreements in our sample of conflicts feature these provisions.

A Provision Selection

Table A.1 lists all provisions in the data and citations for their positive effect on agreement duration. This list represents the pool of candidate indicators for inclusion in our measurement model, but not all provisions are employed. We discuss this process at length in the section on agreement strength measurement.

To assess which indicators are suitable for inclusion in our measurement model, we plot the densities of the indicator discrimination parameters. If all indicators have a positive effect on the latent quantity, then the densities of all parameters should be well to the right of zero. As the parameters are constrained to be positive, no densities will be to the left of

zero, but an indicator that does not have a positive relationship with the latent quantity will have a density that bumps right against zero. Any indicator that is concentrated against zero may be representative of a different latent quantity than that represented by indicators with γ values greater than 0.

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

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

B Descriptive Statistics

Table B.1 presents the proportions of the observed indicators used in our analysis, while Table B.2 breaks them down by peace agreement comprehensiveness. There is significant variation across agreement comprehensiveness, and some indicators such as power sharing or civil service integration are never present in process agreements. The proportion of many indicators coded as 1 decreases as we move from full, to partial, to process peace agreements. These differences support our argument that there are qualitative differences between peace agreements with different levels of comprehensivenss, and that a model which tries to measure

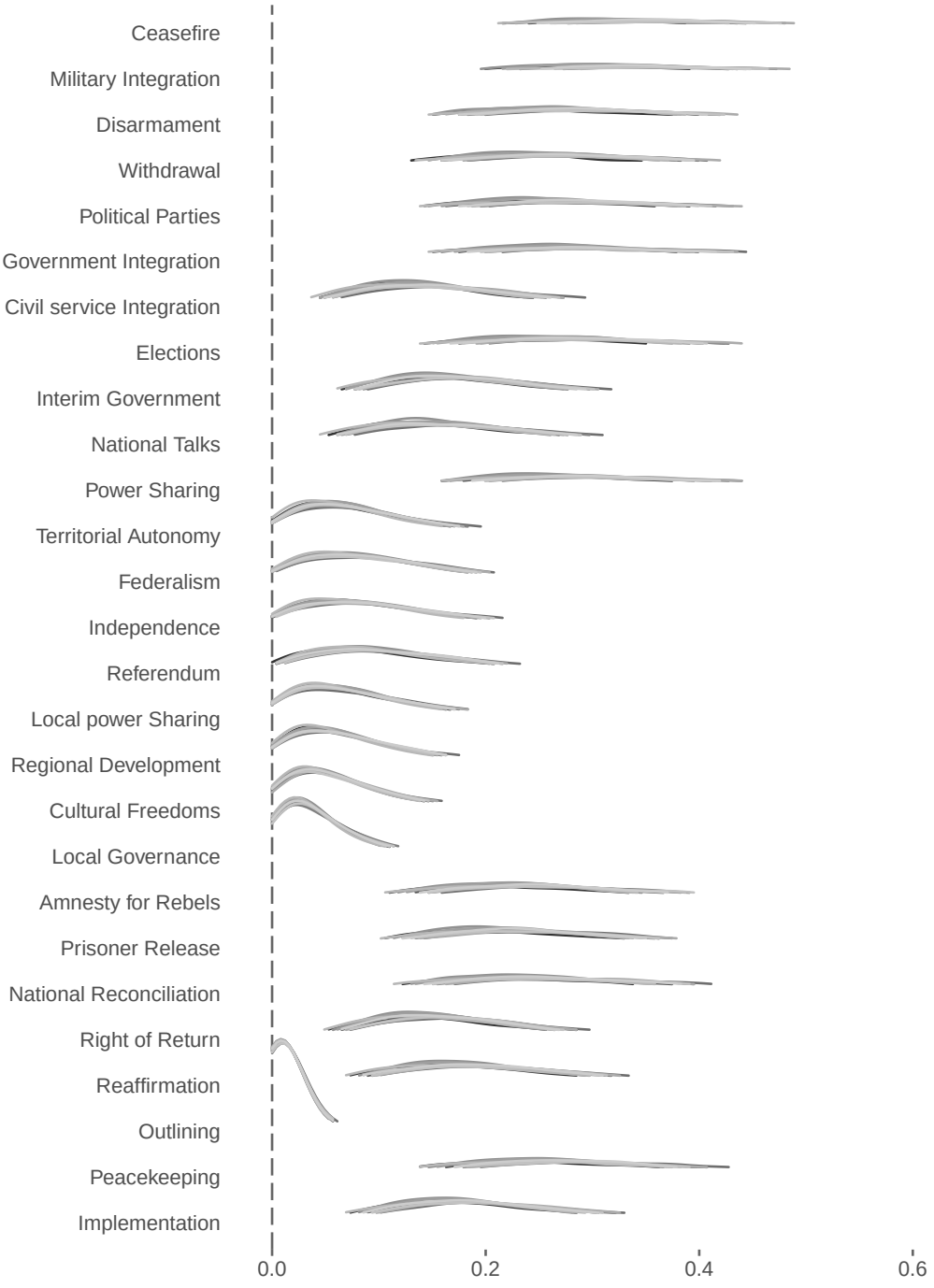


Figure A.1: Discrimination parameters for all agreement provisions. Parameters with the majority of their density near zero represent a different latent dimension. We remove these parameters from our analysis

the strength of all of them pooled together may reach biased conclusions.

	0	1
Ceasefire	0.38	0.62
Military Integration	0.62	0.38
Disarmament	0.65	0.35
Withdrawal	0.85	0.15
Political Parties	0.80	0.20
Government Integration	0.85	0.15
Civil service Integration	0.88	0.12
Elections	0.63	0.37
Interim Government	0.74	0.26
National Talks	0.90	0.10
Power Sharing	0.86	0.14
Amnesty for Rebels	0.73	0.27
Prisoner Release	0.68	0.32
National Reconciliation	0.78	0.22
Right of Return	0.71	0.29
Reaffirmation	0.80	0.20
Peacekeeping	0.77	0.23
Implementation	0.65	0.35

Table B.1: Agreement Indicator Proportions

Table B.3 presents summary statistics for the uncscaled continuous predictors in the full probability model, while Table B.4 presents the proportions for the discrete predictors. We omit the provision "border demarcation" because it refers to international borders, and hence does not appear in our sample of intrastate conflicts. Figure B.1 displays the distributions of these predictors in graphical form. Figure B.2 presents a correlation plot for the (scaled) explanatory predictors of agreement strength. Values for the continuous predictors represent the average of 5 imputations of any variables with missing data. The median conflict has 2 agreements, and ignoring this structure in our data would bias our estimates, which is why we include a random intercept $\boldsymbol{\delta}$ by conflict.

Figure B.3 presents a scatterplot between the additive index and full probability model estimates of strength for peace agreements in our sample.

	0	1	0	1	0	1
Ceasefire	0.19	0.81	0.52	0.48	0.36	0.64
Military Integration	0.44	0.56	0.62	0.38	0.88	0.12
Disarmament	0.36	0.64	0.78	0.22	0.80	0.20
Withdrawal	0.69	0.31	0.92	0.08	0.92	0.08
Political Parties	0.67	0.33	0.80	0.20	1.00	0.00
Government Integration	0.75	0.25	0.84	0.16	1.00	0.00
Civil service Integration	0.83	0.17	0.86	0.14	1.00	0.00
Elections	0.42	0.58	0.68	0.32	0.84	0.16
Interim Government	0.69	0.31	0.76	0.24	0.76	0.24
National Talks	0.97	0.03	0.84	0.16	0.92	0.08
Power Sharing	0.78	0.22	0.84	0.16	1.00	0.00
Amnesty for Rebels	0.50	0.50	0.78	0.22	0.96	0.04
Prisoner Release	0.44	0.56	0.78	0.22	0.80	0.20
National Reconciliation	0.67	0.33	0.82	0.18	0.88	0.12
Right of Return	0.61	0.39	0.70	0.30	0.88	0.12
Reaffirmation	0.72	0.28	0.80	0.20	0.92	0.08
Peacekeeping	0.61	0.39	0.86	0.14	0.84	0.16
Implementation	0.47	0.53	0.72	0.28	0.76	0.24
(a) Full			(b) Partial		(c) Process	

Table B.2: Observed Indicator Proportions by Agreement Type

	min	max	mean	median	sd
Aid (% GNI)	0.00	0.52	0.10	0.07	0.11
RPR	0.29	1.61	1.03	1.09	0.32
Polity2	-9.00	10.00	0.58	0.00	5.05

Table B.3: Summary statistics for continuous predictors. Values reflect the average of five imputed datasets due to missingness on all continuous predictors. Predictors are centered and scaled in models, but shown untransformed.

	0	1	2
Sanction	0.65	0.08	0.27
Mediation	0.49	0.45	0.06
Intervention (IMI)	0.48	0.52	
Intervention (ACD)	0.86	0.14	
Comprehensiveness	0.32	0.45	0.23
Government	0.20	0.80	
Cumulative Intensity	0.15	0.85	
Post Cold War	0.12	0.88	

Table B.4: Summary statistics for continuous predictors. Values of two for sanction and mediation represent multilateral sanctions and mediation by regional organizations, respectively. Comprehensiveness is reverse coded so increasing values represent less comprehensive agreements.

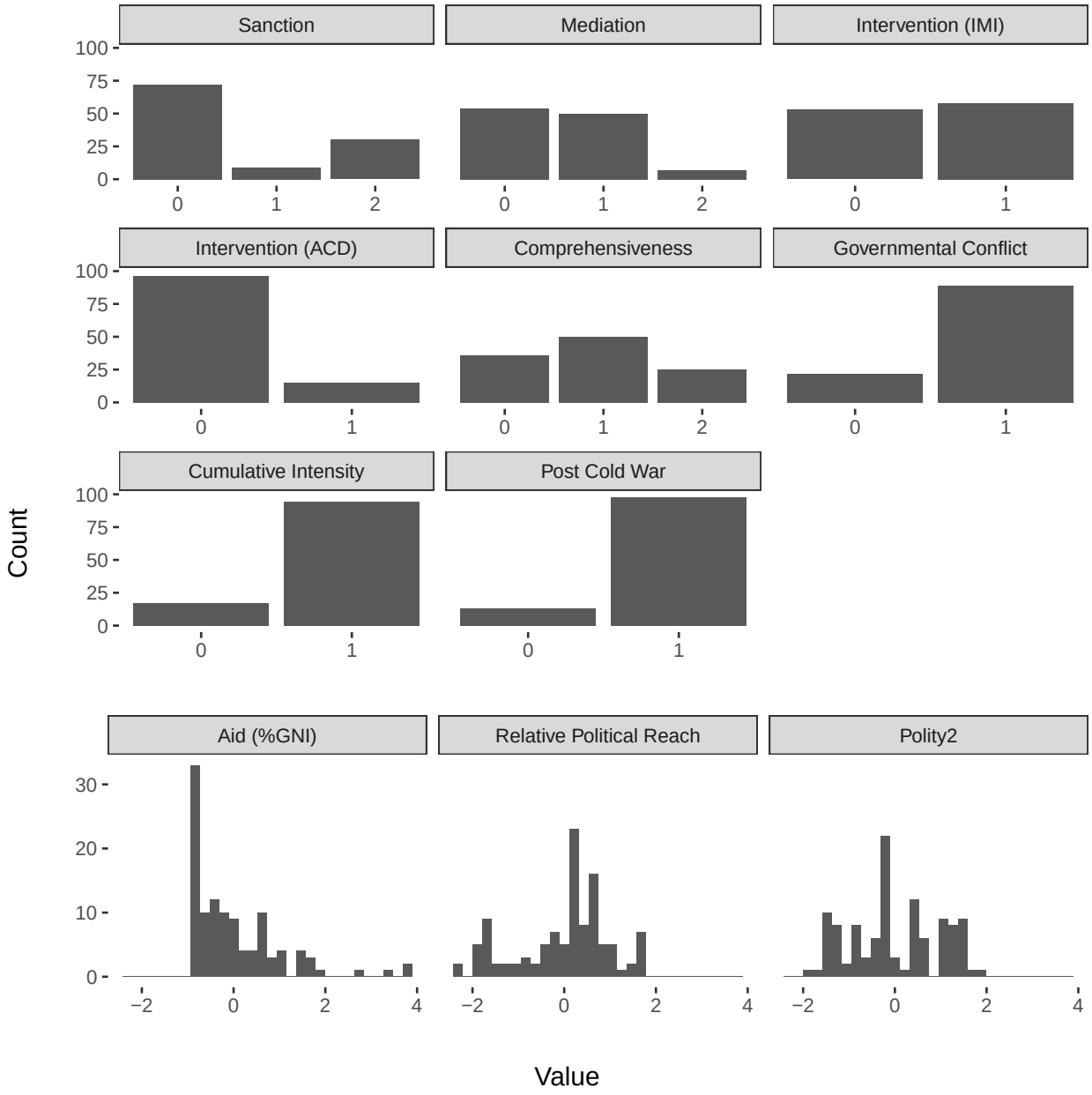


Figure B.1: Distributions of predictors in regression model explaining agreement strength. Continuous predictors are shown centered and standardized

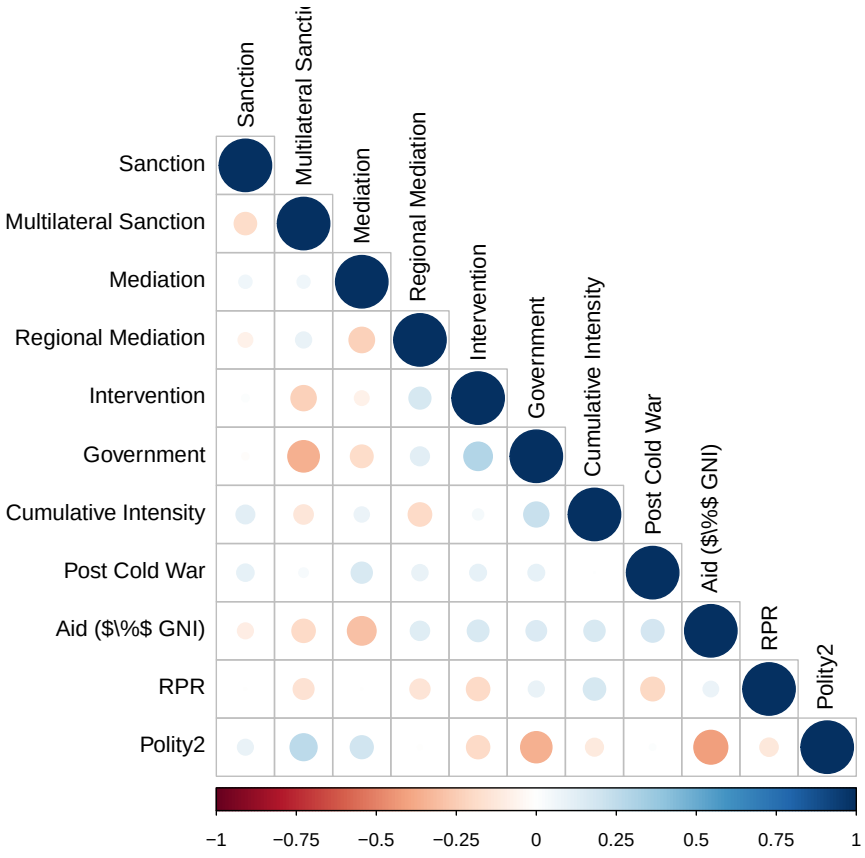


Figure B.2: Correlation of predictors. Continuous predictors centered and standardized.

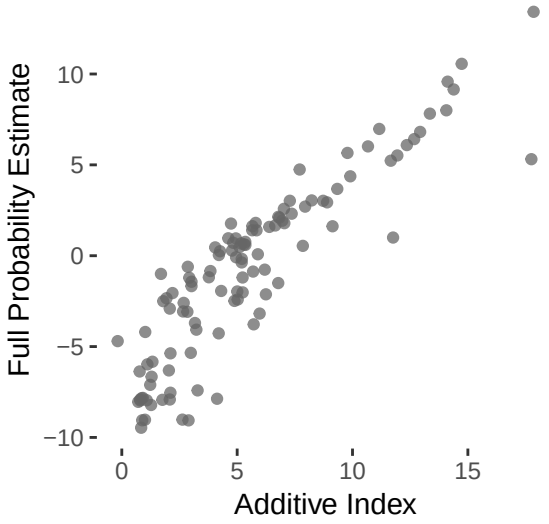


Figure B.3: Scatterplot of latent variable estimate of agreement strength from the full probability model and an additive index of provisions.

C Data Cleaning and Recoding Decisions

There are several observations in the Civil War Mediation data that are missing either start or end dates to the mediation episode. The mean duration of mediation events with no missing start or end dates (n = 461) is 93 days. However, the data are *extremely* skewed with a maxmimum mediation duration of 5,205 days, followed by a duration of 2,477 days. As such, we use the median mediation duration of 6 days and set missing start dates 6 days earlier than their corresponding end dates, and missing end dates 6 days later than their corresponding event dates. Median imputation allows us to include more mediation events than listwise deletion otherwise would, and should not significantly bias our results because it reduces variation in a variable used to code a dummy variable. We also potentially over-count mediation as the available data do not allow us to match on conflict, just state.

The International Military Intervention data feature similar missingness of start and end dates. Any observations with neither a start or end date are dropped. Observations missing only one data follow the same procdure as above. The distribution of intervention durations is similarly skewed with maximum duration of 14,824 days, mean of 836, and median of 155. As with mediation, we set missing start and end dates 155 days before or after the observed date.

The date ranges for each of our data sources are listed below. The start of the Peace Agreements Data in 1975 and the end of the TIES and IMI data in 2005 dictate our sample of agreements from 1975-2005.

- UCDP Armed Conflict Data[1]: 1946-2012
- UCDP Peace Agreements: 1975-2011
- Civil War Mediation: 1946-2011
- Threat and Imposition of Economic Sanctions: 1945-2005
- Relative Political Capacity: 1960-2013

[1]We use Version 4-2013 of the ACD for compatibility with the ID system used by the Peace Agreements Data

- World Development Indicators: 1960-2017
- Polity: 1800-2015
- International Military Interventions: 1946-2005

D Missing Data and Multiple Imputation

Variables in the analysis dataset with missing values, and the proportion of missingness are presented in Table D.1. As the standard practice is to only impute variables with fewer than 15% missing values, we multiply impute missing values for all of these variables, generating 5 imputed datasets.

	% Missing
Aid (% GNI)	12.61
RPR	6.31
Polity2	1.80

Table D.1: Variables with Missing Values

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

Varying Anchoring Agreements

The distribution of agreements along this latent scale is relatively invariant to different choices of agreements for the strong and weak identification restriction. In addition to the results from the model above, Figure D.1 presents the distribution of agreement strength for models with three different sets of identification restrictions. The position of the points in Figure D.1 represents how far an agreement has moved in the order compared to the model in Figure **??**. The x axis is the order in the main model whose results are presented here, while the y axis represents the order in models with three different choices of weak and strong agreements to identify the scale. Points exactly on the diagonal indicate an agreement whose position in the ranking of agreements is identical under both sets of identification restrictions.

An unstable measure would see few points near the diagonal, while a stable one would see many points along the diagonal. Most of the points in Figure D.1 are relatively close to the diagonal. This suggests the latent scale of agreement strength is not sensitive to choice of identification restriction. We present the coefficients for agreement strength predictors from each of these models in the Online Appendix; the magnitude and direction of estimates is stable with regard to choice of identification restriction.

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

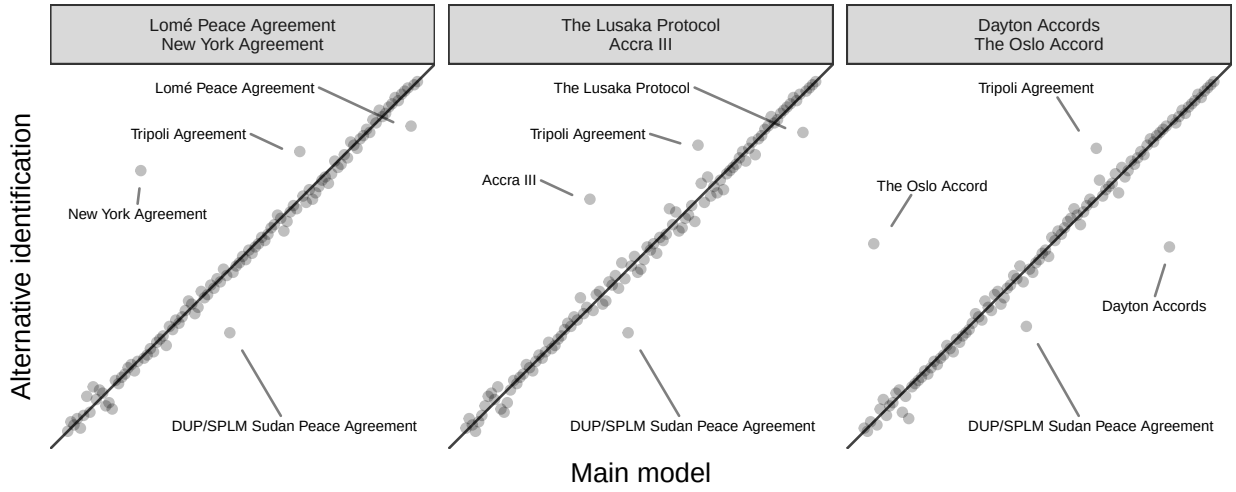


Figure D.1: Comparison of agreement strength estimates from models with three different sets of end points selected as identification restrictions. Points below the diagonal indicate agreements that are ranked higher in our main model, while those above the diagonal indicate agreements ranked higher under an alternative identification strategy. Agreements which shift more than 10 places in the ranking are labeled.

especially policy relevant agreements for identification restrictions.

E Additional Results

The units of scale for agreement strength are not inherently meaningful because they are the result of a latent variable estimation. Accordingly, they should be thought of relative to the total extent of agreement strength values. Interpreting the relationship between peace agreement strength and both types of sanctions and mediation along with military intervention is relatively straightforward because they are dummy or factor variables, so their marginal effect is just the result of their presence or absence relative to the excluded category. The effect of aid is less straightforward, but relatively simple, because it is standardized to have mean 0 and unit variance. Figure E.1 presents the median effect of a one unit shift in

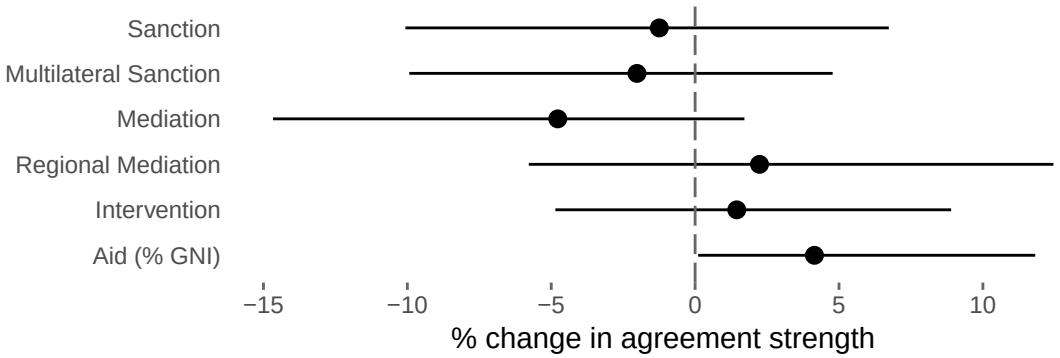


Figure E.1: Estimates of the marginal effect of a one unit shift in predictors on agreement strength, with 95% credible intervals. A 5% change in agreement strength means that a dummy variable's presence moves the estimate 5% up the range of agreement strength, or that a one unit change in a continuous variable's value moves the estimate the same distance.

our predictors on an agreement's position within the range of agreement strength estimates. The median effect of an agreement being signed under the duress of economic sanctions is -0.28, which shifts an agreement's strength 1.24% downward in the distribution of agreement strengths, which is not as far down as the -0.46 median effect of multilateral sanctions which shifts agreements 2.02% downward. Foreign aid, the only predictor with more than 95% of its posterior distribution on the same side of zero, produces a change of 0.95 for each one unit shift in scaled aid, which translates to a 4.15% shift upward in the distribution of agreement strengths.

F Full Probability Model vs. Standalone IRT

To demonstrate that our results are not an artifact of model specification choices, we present a comparison of our results along with those taken from a model which estimates the latent agreement strength and the effect of our explanatory variables on agreement strength separately. We also present a model which uses only one of our explanatory variables – sanctions – and show that the results of this bivariate model are similar to those of our full model.

Figures F.1 presents the distribution of estimated agreement strengths, and their uncertainties, from the full probability model and the standalone IRT model. The estimates without any associated uncertainty are the two fixed agreements used to identify and scale our model. The full probability model estimates are taken from the model that incorporates all explanatory predictors. Since there is no missingness in the observed indicators, we do not need to perform any imputation and run 4 chains for our standalone IRT model.

The colors of the points range from darkest to lighest representing their position within the spectrum of agreement strengths in the full probability model. Thus, a point whose color is out of step with its neighbors' in the lower plot represents an agreement whose place in the ranking shifts significantly between the two models. While there are some points in the middle of the lower which have shifted position from the full probability model, those at the extremes remain relatively consistent, telling us that their are not fundamental differences between the two estimated scales. These minor differences can be explained by the fact that the full probability model includes more information than the standalone IRT model.

Differences in ordering between the two estimates represent increased accuracy in the full probability model due to the extra information it is able to incorporate. This is especially true at the bottom end of the scale where several agreements have identical strengths in the standalone IRT mode. The higher level of information in the full probability model introduces more variation into the estimates, allowing for better inference.

Figure F.2 presents a scatter plot of relative agreement strengths between the full probability model and the standalone IRT model. In contrast with the comparison of different

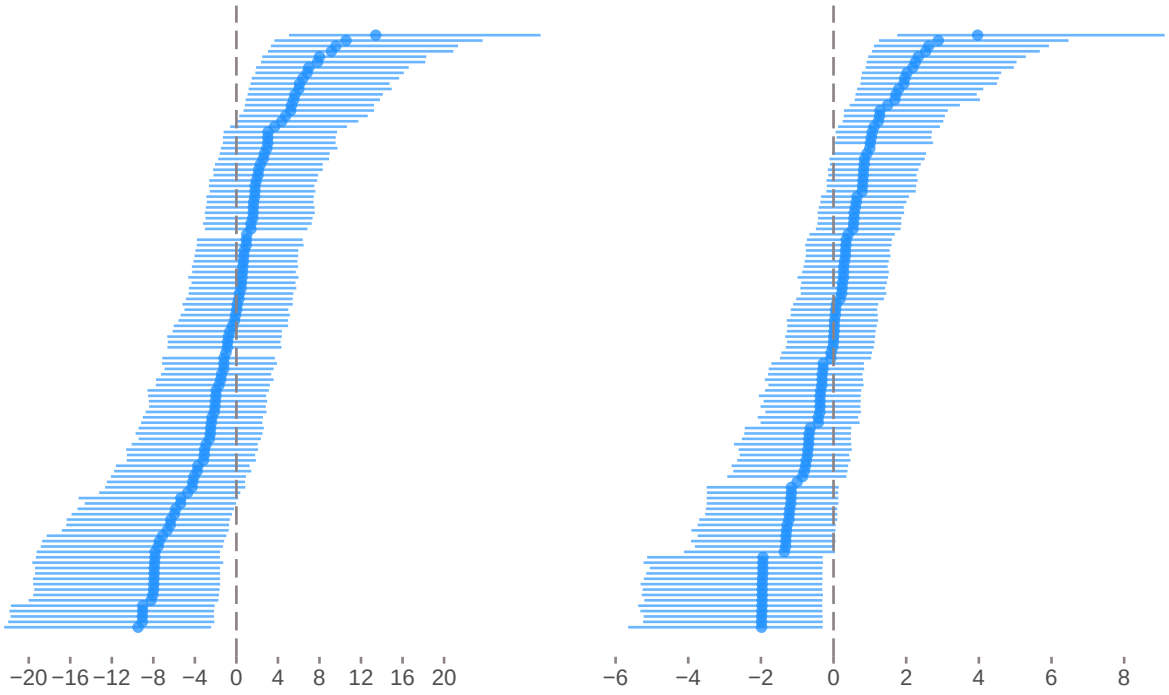


Figure F.1: Latent Agreement Strengths

identification strategies under the full probability model in the paper, many agreements move more than 10 positions in the rankings. While the overall distribution of agreement strengths is relatively similar, as seen in Figure F.2, the position of agreements within those distributions varies significantly, with 12 moving more than 10 places in ranking.

We first present a series of scatter plots from our main model. Figure F.3 shows the distibution of (jittered) agreement strengths relative to our explanatory variables, along with best fit lines. Although the comprehensiveness of a peace agreement is a control variable, we briefly discuss results for it as a way to verify that our latent scale is correctly estimated.

Figure F.4 presents the distibution of (jittered) agreement strengths from the standalone IRT model relative to our explanatory variables, along with best fit lines. Although the estimated agreement strengths are different from those produced by the full probability model in Figure F.3, the relationships between them and our explanatory variables are substantively similar.

Table F.1 presents the estimates of the effect of the explanatory variables on agreement

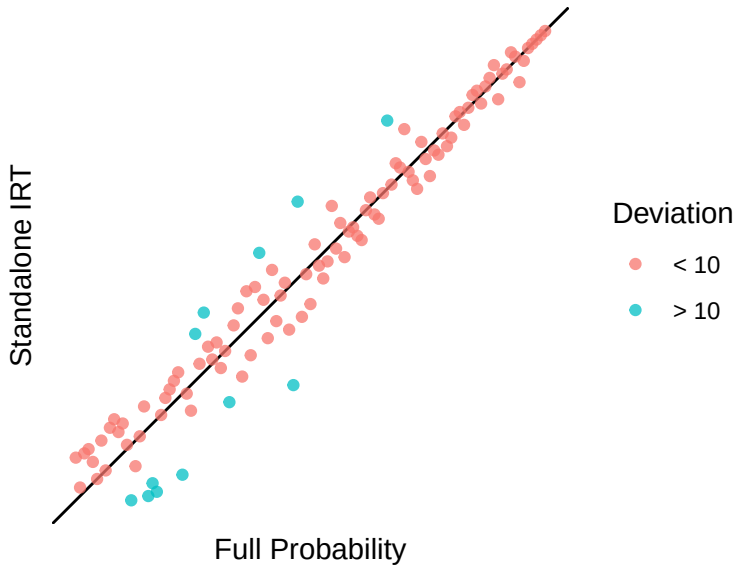


Figure F.2: Comparison of agreement strength estimates from the full probability and standalone IRT models. Points below the diagonal indicate agreements that are ranked higher in the former, while those above the diagonal indicate agreements ranked higher in the latter. Agreements which shift more than 10 places in the ranking are labelled.

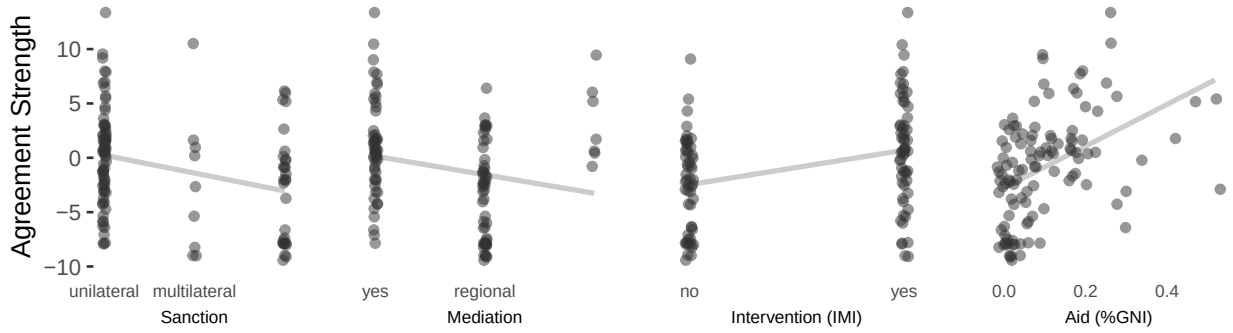


Figure F.3: Scatter plots of (jittered) agreement strength and explanatory variables.

strength in the full probability model, with and without the other explanatory predictors, and in the separate IRT and linear model, also with and without the other explanatory predictors. Including all predictors in both the full probability and standalone IRT models shifts the posterior mean slightly upward, but in both cases, the majority of the posterior

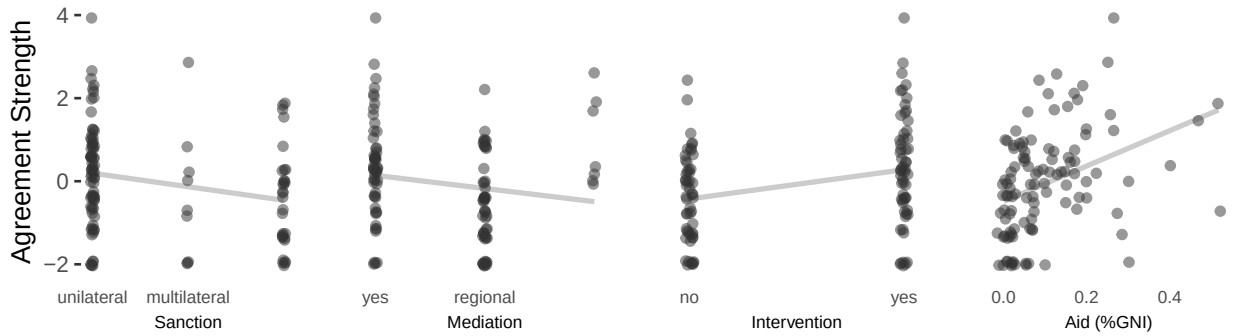


Figure F.4: Bivariate Scatter Plots from Standalone IRT Model

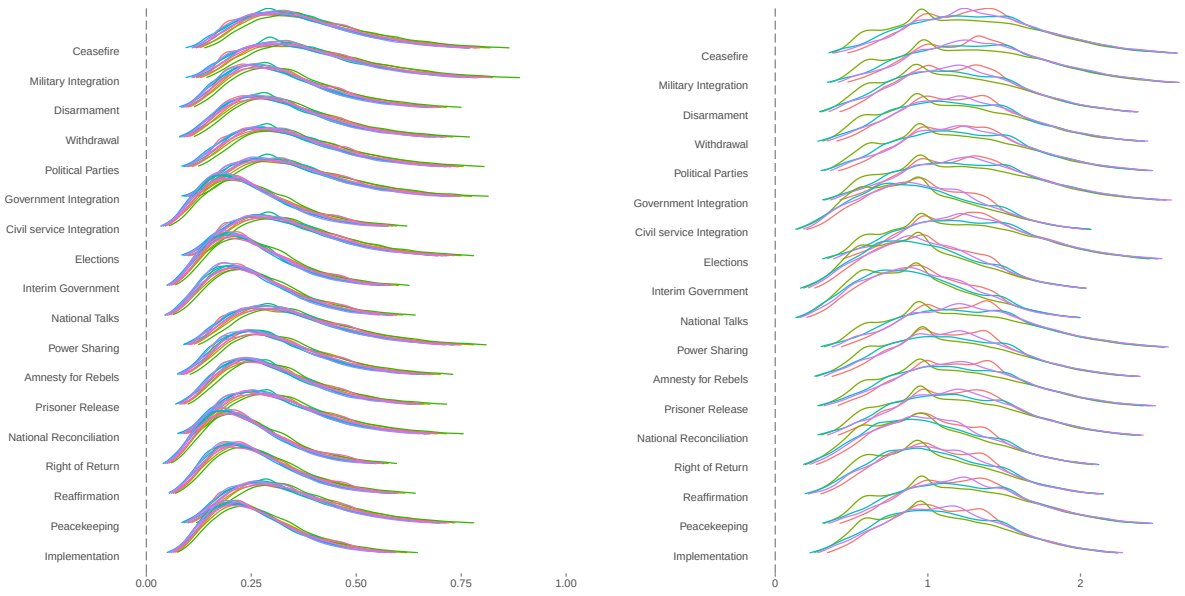
distribution is still less than zero. This suggests that the negative effect of sanctions on agreement strength persists even when controlling for potentially confounding sources of variation.

Figures F.5a and F.5b present the distributions of the discrimination parameters for the full probability model, and the standalone IRT model, respectively. There is more distance between the distributions and zero in the standalone IRT model, but the distributions are much wider, indicating greater uncertainty about the parameters. There is also less overlap between the chains, which means that the sampler has had more difficulty converging to the stationary distribution. The extra information included in the full probability model yields less uncertain estimates and results in more efficient sampling from the posterior.

	Standalone IRT	Standalone IRT	Full Probability	Full Probability
Sanction	−0.25	−0.41	−0.27	−0.31
	[−1.00; 0.50]	[−1.17; 0.34]	[−1.98; 1.33]	[−2.30; 1.54]
Multilateral Sanction	−0.17	−0.29	−0.48	−0.50
	[−0.72; 0.39]	[−0.86; 0.29]	[−2.20; 0.87]	[−2.27; 1.09]
Mediation	−0.34	−0.46	−0.93	−1.19
	[−0.83; 0.14]	[−0.96; 0.02]	[−2.66; 0.34]	[−3.36; 0.39]
Regional Mediation	0.50	0.51	0.55	0.58
	[−0.37; 1.36]	[−0.35; 1.38]	[−1.06; 2.50]	[−1.32; 2.85]
Intervention	0.25	0.13	0.43	0.37
	[−0.23; 0.72]	[−0.37; 0.62]	[−0.81; 1.95]	[−1.11; 2.04]
Aid (% GNI)	0.33*	0.19	0.97*	1.06*
	[0.08; 0.59]	[−0.08; 0.46]	[0.16; 2.35]	[0.02; 2.71]
μ_δ	0.15	−0.53	0.00	−1.32
	[−0.35; 0.65]	[−1.52; 0.45]	[−2.94; 2.78]	[−5.49; 2.86]
Government		−0.26		0.27
		[−1.05; 0.52]		[−1.45; 2.27]
Cumulative Intensity		0.74*		0.91
		[0.02; 1.49]		[−0.89; 3.18]
Post Cold War		0.51		0.62
		[−0.16; 1.17]		[−1.12; 2.66]
RPR		−0.18		−0.64
		[−0.54; 0.17]		[−2.05; 0.52]
Polity2		−0.25		−0.52
		[−0.52; 0.03]		[−1.88; 0.59]

* 0 outside 95% credible interval

Table F.1: Full Probability Model and Standalone IRT



(a) Full Probability Model (b) Standalone IRT

Figure F.5: Discrimination Parameters

G Robustness Checks

	α	γ
Ceasefire	-0.85	0.40
Military Integration	0.58	0.40
Disarmament	0.70	0.34
Withdrawal	2.07	0.34
Political Parties	1.74	0.36
Government Integration	2.12	0.37
Civil service Integration	2.22	0.26
Elections	0.61	0.35
Interim Government	1.17	0.27
National Talks	2.42	0.27
Power Sharing	2.20	0.37
Amnesty for Rebels	1.17	0.33
Prisoner Release	0.84	0.31
National Reconciliation	1.55	0.33
Right of Return	0.99	0.25
Reaffirmation	1.59	0.28
Peacekeeping	1.52	0.35
Implementation	0.66	0.27

Table G.1: Difficulty (α) and discrimination (γ) parameters in the measurement model. The difficulty parameter controls the location of the item characteristic curve's inflection point, while the discrimination parameter controls the slope.

Table G.1 presents the estimated difficulty and discrimination parameters, which are presented graphically in the paper, in numeric form.

Table G.2 presents the results from Model 6 in the paper with 95% highest posterior density intervals (HPDI) instead of credible intervals. Using HPDI instead of equal tailed credible intervals accounts for asymmetric posterior distributions and presents an interval that captures the highest 95% posterior density. The interval bounds are similar to the credible intervals presented in the paper, but the 95% HPDI interval for foreign aid now contains zero, suggesting that there is greater uncertainty about its effect.

The `pa_type` variable in the peace agreements data list whether an agreement is a process, partial, or full agreement. The codebook (Harbom, Högbladh & Wallensteen 2006) describes each type of agreement as follows:

	Full Probability Model
Sanction	−0.31
	[−2.27; 1.57]
Multilateral Sanction	−0.50
	[−2.21; 1.15]
Mediation	−1.19
	[−3.13; 0.54]
Regional Mediation	0.58
	[−1.43; 2.72]
Intervention	0.37
	[−1.17; 1.96]
Aid (% GNI)	1.06
	[−0.10; 2.47]
Government	0.27
	[−1.51; 2.19]
Cumulative Intensity	0.91
	[−1.03; 3.00]
Post Cold War	0.62
	[−1.19; 2.57]
RPR	−0.64
	[−1.97; 0.58]
Polity2	−0.52
	[−1.82; 0.64]
μ_δ	−1.32
	[−5.52; 2.83]

* 0 outside 95% highest posterior density interval

Table G.2: Posterior Density of Conflict Level Predictors

	Main Model	Comprehensiveness	No Peacekeeping
Sanction	−0.31	−0.29	−0.29
	[−2.30; 1.54]	[−1.85; 1.21]	[−2.16; 1.49]
Multilateral Sanction	−0.50	−0.68	−0.47
	[−2.27; 1.09]	[−2.09; 0.57]	[−2.18; 1.07]
Mediation	−1.19	−1.08	−1.26
	[−3.36; 0.39]	[−2.60; 0.09]	[−3.38; 0.25]
Regional Mediation	0.58	0.23	0.58
	[−1.32; 2.85]	[−1.23; 1.83]	[−1.20; 2.69]
Intervention	0.37	0.08	0.22
	[−1.11; 2.04]	[−1.09; 1.29]	[−1.23; 1.74]
Aid (% GNI)	1.06*	0.49	0.92
	[0.02; 2.71]	[−0.18; 1.39]	[−0.03; 2.43]
Government	0.27	0.18	0.33
	[−1.45; 2.27]	[−1.25; 1.78]	[−1.37; 2.27]
Cumulative Intensity	0.91	0.90	1.01
	[−0.89; 3.18]	[−0.56; 2.61]	[−0.82; 3.23]
Post Cold War	0.62	0.93	0.74
	[−1.12; 2.66]	[−0.53; 2.69]	[−0.98; 2.71]
RPR	−0.64	−0.42	−0.57
	[−2.05; 0.52]	[−1.40; 0.39]	[−1.92; 0.54]
Polity2	−0.52	−0.52	−0.47
	[−1.88; 0.59]	[−1.48; 0.24]	[−1.76; 0.55]
Partial		−1.32*	
		[−2.80; −0.17]	
Process		−3.30*	
		[−6.27; −1.46]	
μ_δ	−1.32	−0.02	−1.65
	[−5.49; 2.86]	[−3.07; 3.13]	[−5.69; 2.33]

* 0 outside 95% credible interval

Table G.3: Statistical Models

- A full agreement is an agreement where one or more dyad agrees to settle the whole in- compatibility.
- A partial peace agreement is an agreement where one or more dyad agrees to settle a part of the incompatibility.
- A peace process agreement is an agreement where one or more dyad agrees to initiate a process that aims to settle the incompatibility.

Table G.3 presents results from models that include agreement comprehensiveness as an explanatory variable and that exclude the peacekeeping as a provision. Process agreements are weaker than partial ones, which are weaker from comprehensive ones. This relationship suggests that our latent scale is oriented correctly, but also raises endogeneity concerns. If comprehensiveness is also a measure of agreement strength, then including it as an explanatory variable could bias our estimates. For this reason, we exclude it from the models in the paper. Implementation denotes whether an "agreement provided for the establishment of a commission or committee to oversee implementation of the agreement" and peacekeeping indicates whether an "agreement provided for the deployment of a peace-keeping operation" (Harbom, Högbladh & Wallensteen 2006). Neither reflects whether an agreement was actually implemented or if peacekeeping actually occurred, so they do not introduce post-treatment bias. However, peacekeeping could indicate a level of third-party interest in the conflict, raising concerns that international factors are used both to measure and explain agreement strength. As the final model in Table G.3 shows, omitting peacekeeping as an indicator does not substantively change any parameter estimates for our explanatory variables.

Table G.4 presents results using the the ACD `type` variable which discriminates between internal armed conflict and *internationalized* internal armed conflict which denotes external military intervention into a conflict in a given year. The estimates for intervention have opposite signs depending on which measure is used, but estimates for all other predictors are

	IMI	ACD	IMI	ACD
Sanction	−0.27	−0.22	−0.31	−0.33
	[−1.98; 1.33]	[−1.97; 1.38]	[−2.30; 1.54]	[−2.22; 1.46]
Multilateral Sanction	−0.48	−0.48	−0.50	−0.52
	[−2.20; 0.87]	[−2.03; 0.85]	[−2.27; 1.09]	[−2.25; 1.01]
Mediation	−0.93	−1.00	−1.19	−1.15
	[−2.66; 0.34]	[−3.41; 0.36]	[−3.36; 0.39]	[−3.17; 0.39]
Regional Mediation	0.55	0.58	0.58	0.58
	[−1.06; 2.50]	[−1.60; 2.63]	[−1.32; 2.85]	[−1.37; 2.83]
Intervention	0.43	−0.39	0.37	−0.58
	[−0.81; 1.95]	[−2.09; 1.17]	[−1.11; 2.04]	[−2.52; 1.21]
Aid (% GNI)	0.97*	1.03*	1.06*	1.03*
	[0.16; 2.35]	[0.18; 2.40]	[0.02; 2.71]	[0.03; 2.80]
μ_δ	0.00	0.19	−1.32	−1.10
	[−2.94; 2.78]	[−3.15; 2.98]	[−5.49; 2.86]	[−5.38; 2.82]
Government			0.27	0.28
			[−1.45; 2.27]	[−1.42; 2.31]
Cumulative Intensity			0.91	1.04
			[−0.89; 3.18]	[−0.81; 3.39]
Post Cold War			0.62	0.55
			[−1.12; 2.66]	[−1.14; 2.51]
RPR			−0.64	−0.67
			[−2.05; 0.52]	[−2.03; 0.50]
Polity2			−0.52	−0.54
			[−1.88; 0.59]	[−1.88; 0.54]

* 0 outside 95% credible interval

Table G.4

unaffected. The correlation between the intervention variables is 0.22, which reflects the fact that the ACD internationalized conflict variable is a much less nuanced measure of military intervention.

Table G.5 presents coefficient estimates for the explanations of agreement strength for our main model (Model 1) and the three differently identified models presented in the paper (Models 2-4). While a few predictors gain or lose significance, the magnitude and direction of estimates is stable across specifications. This robustness provides further evidence that our findings are not an artifact of our choice of which agreements we use as identification restrictions.

	Model 1	Model 2	Model 3	Model 4
Sanction	−0.31	−0.35	−0.40	−0.15
	[−2.30; 1.54]	[−1.91; 1.14]	[−2.03; 1.26]	[−2.24; 1.85]
Multilateral Sanction	−0.50	−0.40	−0.22	−0.61
	[−2.27; 1.09]	[−1.76; 0.79]	[−1.71; 1.12]	[−2.49; 1.31]
Mediation	−1.19	−0.87	−0.72	−1.14
	[−3.36; 0.39]	[−2.64; 0.22]	[−2.29; 0.50]	[−3.38; 0.54]
Regional Mediation	0.58	0.41	0.52	0.58
	[−1.32; 2.85]	[−1.16; 2.14]	[−0.88; 2.48]	[−1.35; 2.98]
Intervention	0.37	0.22	0.30	0.44
	[−1.11; 2.04]	[−1.06; 1.50]	[−0.91; 1.60]	[−1.18; 2.39]
Government	0.27	−0.09	0.08	0.20
	[−1.45; 2.27]	[−1.68; 1.38]	[−1.47; 1.61]	[−1.67; 2.26]
Cumulative Intensity	0.91	0.73	0.73	0.95
	[−0.89; 3.18]	[−0.66; 2.36]	[−0.95; 2.57]	[−1.09; 3.55]
Post Cold War	0.62	0.52	0.52	0.57
	[−1.12; 2.66]	[−0.86; 1.90]	[−1.02; 2.10]	[−1.17; 2.70]
Aid (% GNI)	1.06*	0.73	0.71	1.12
	[0.02; 2.71]	[−0.01; 2.19]	[−0.11; 1.81]	[−0.13; 2.88]
RPR	−0.64	−0.41	−0.52	−0.72
	[−2.05; 0.52]	[−1.44; 0.40]	[−1.62; 0.40]	[−2.35; 0.60]
Polity2	−0.52	−0.60	−0.79	−1.03
	[−1.88; 0.59]	[−1.62; 0.07]	[−2.25; 0.00]	[−2.76; 0.18]
μ_δ	−1.32	−3.01*	−3.14	−2.18
	[−5.49; 2.86]	[−6.71; −0.31]	[−7.12; 0.03]	[−6.61; 2.30]

* 0 outside 95% credible interval

Table G.5: Posterior density of coefficient estimates for full probability model estimated with four different identification restrictions. Model 1 is the main model using the DUP/SPLM Sudan Peace Agreement and Tripoli Agreement as low and high strength end points, respectively. Model 2 uses the New York Agreement and Lomé Peace Agreement, Model 3 uses Accra III and The Lusaka Protocol, while Model 4 uses The Oslo Accord and the Dayton Accord. The sign of each of our variables of interest are consistent, and their magnitudes are substantively similar. Using the zero exclusion criterion to assign significance, a few variables gain or lose significance depending on the end points used.

H Convergence Diagnostics

Figure H.1 presents the distribution of all $\hat{R}$ statistics in the full probability model. Values below 1.1 are generally considered 'good' (Gelman & Rubin 1992), and all of our $\hat{R}$ statistics are ≤ 1.1, so we have further evidence indicating good exploration of the parameter space.

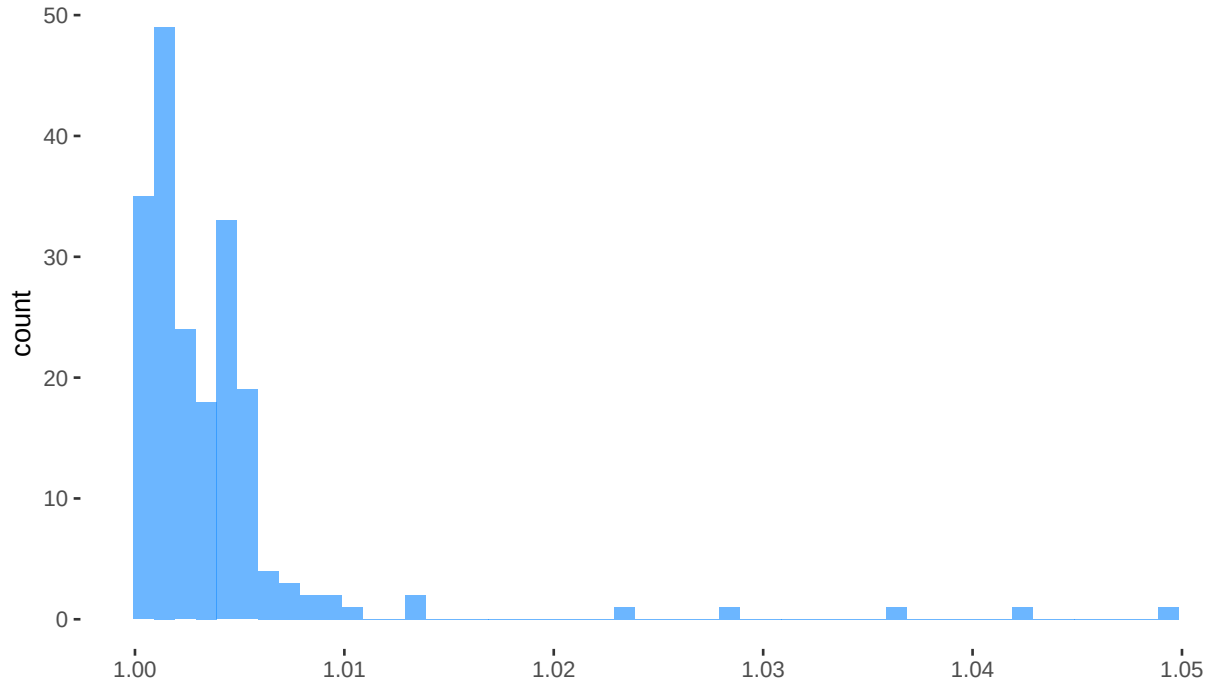


Figure H.1: Histogram of $\hat{R}$ Statistics in Full Probability Model

Table H.1 presents the results of the Heidelberger-Welch diagnostic run on all 10 chains, averaged over all parameters in each chain. The first column presents the proportion of parameters in each chain that passes the stationarity test. The second column presents the average starting interation in each chain used to perform the stationarity test. Values close to 1 indicate that we have evidence that earlier samples for that chain are from the stationary distribution. The third column presents the average p-value for the stationarity test. The test assumes the null hypothesis that the samples are from the stationary distribution, so the lowest p-value of 0.27 means we fail to reject the null, thus providing us with evidence that all 10 have converged to the stationary distribution.

Figure H.2 shows the Z-scores from the Geweke Diagnostic which compares the means in

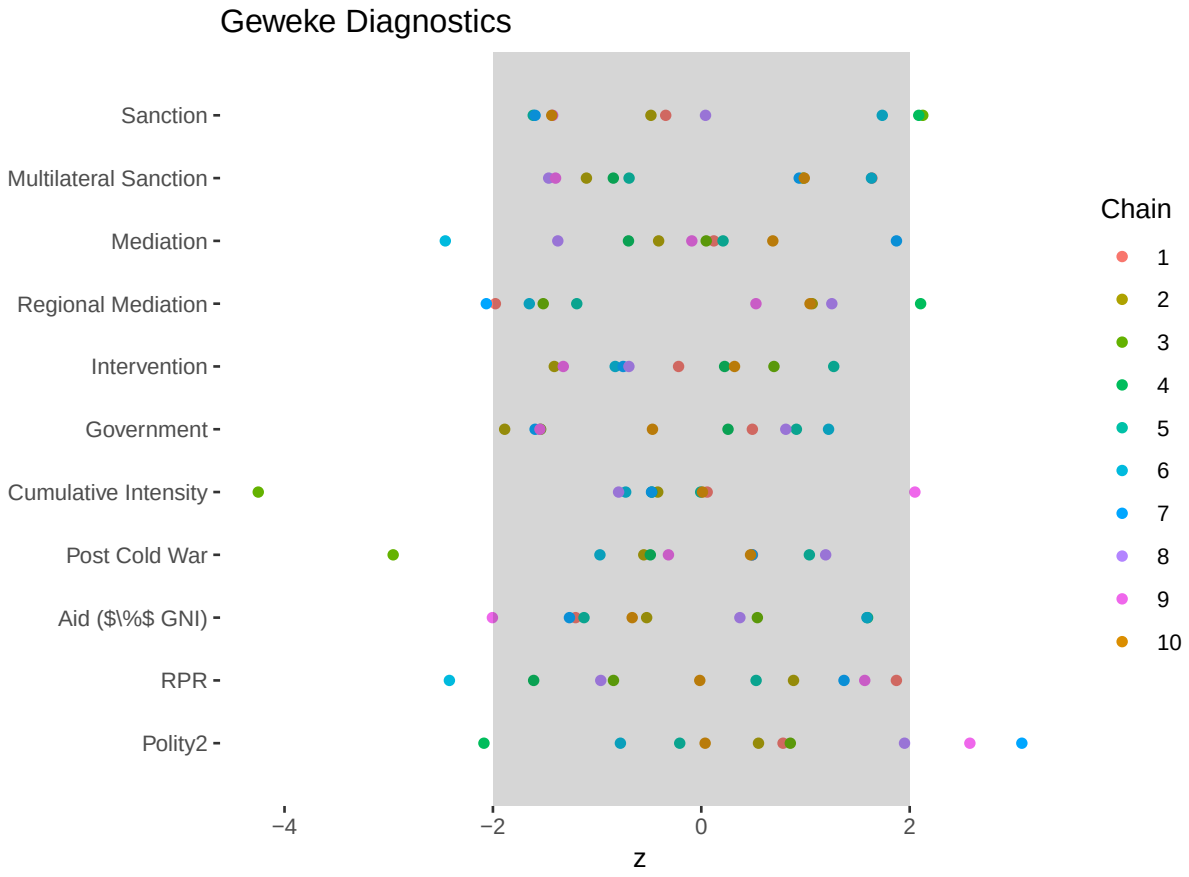


Figure H.2: Geweke Diagnostic Test Statistics

	prop. Passed	Starting Iteration	p-value
1	1.00	124.71	0.43
2	1.00	1.00	0.80
3	1.00	616.38	0.27
4	1.00	521.41	0.36
5	1.00	185.62	0.63
6	1.00	5570.23	0.28
7	1.00	123.45	0.36
8	1.00	31.61	0.51
9	1.00	521.41	0.38
10	1.00	368.35	0.65

Table H.1

two different fractions of each chain (Geweke 1992). We follow standard practice and compare the first 10% of each chain with the final 50%. The diagnostic tests the null hypothesis that both fractions are drawn from the stationary distribution, and we can see from the plot that we fail to reject this hypothesis because almost all Z-scores are within two standard deviations from the mean.

Figures H.4 and H.3 present traceplots for the measurement model parameters and regression model coefficients in the full probability model. Figure H.5 presents the within-chain autocorrelation plots for the measurement model parameters and regression model coefficients in the full probability model, averaged across all 10 chains.

Figure H.5 presents the histograms of autocorrelation within chains, which measures the inefficiency introduced by first order dependence of the Markov process used in sampling. Autocorrelation falls off significantly faster for predictor coefficients β than discrimination parameters γ. The decreased autocorrelation of sampling β improves the overall efficiency of the sampler relative to a standalone IRT model which does not include β.

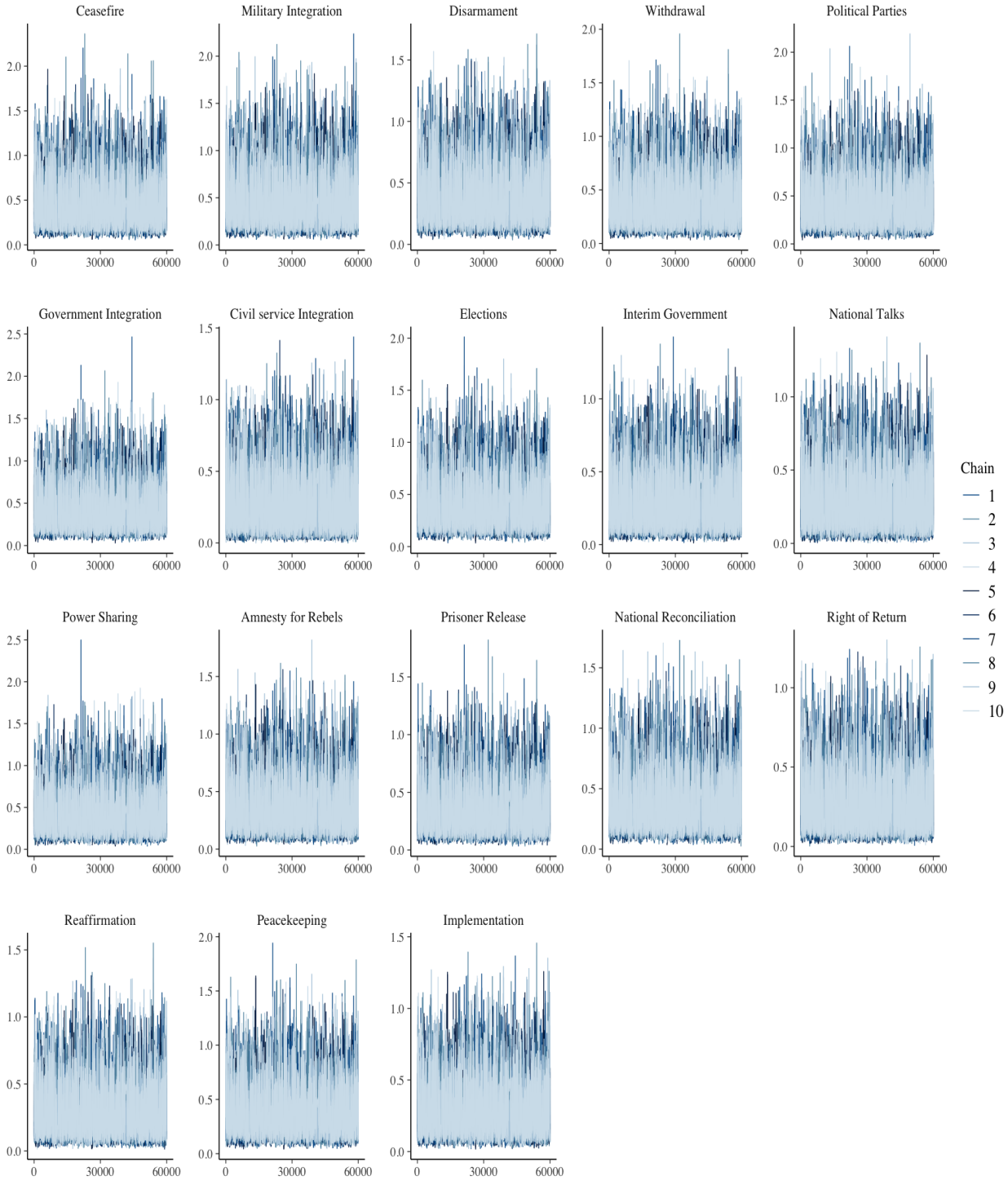


Figure H.3: Traceplot for measurement model discrimination parameters in the full probability model. The overlap indicates good mixing between chains.

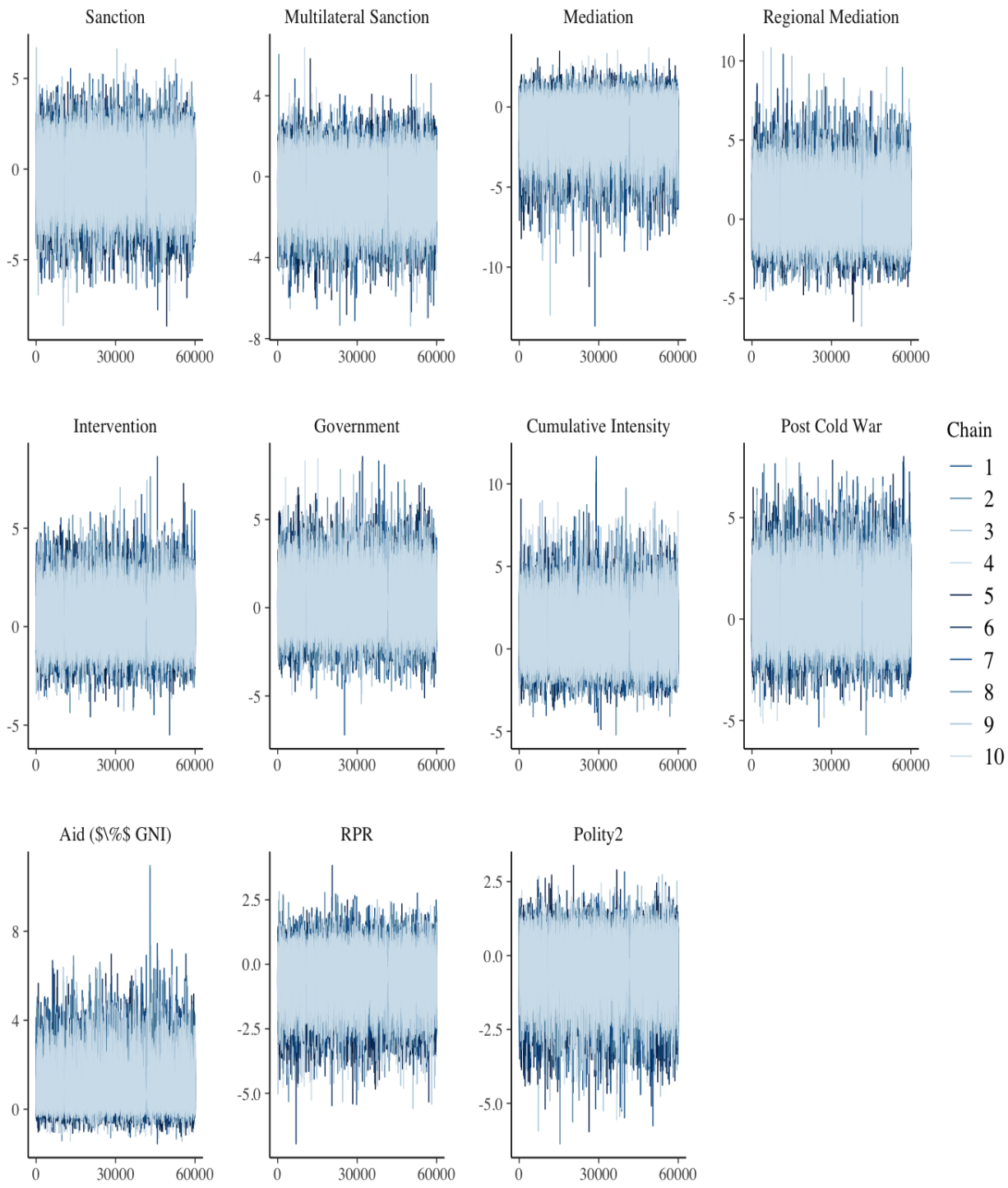
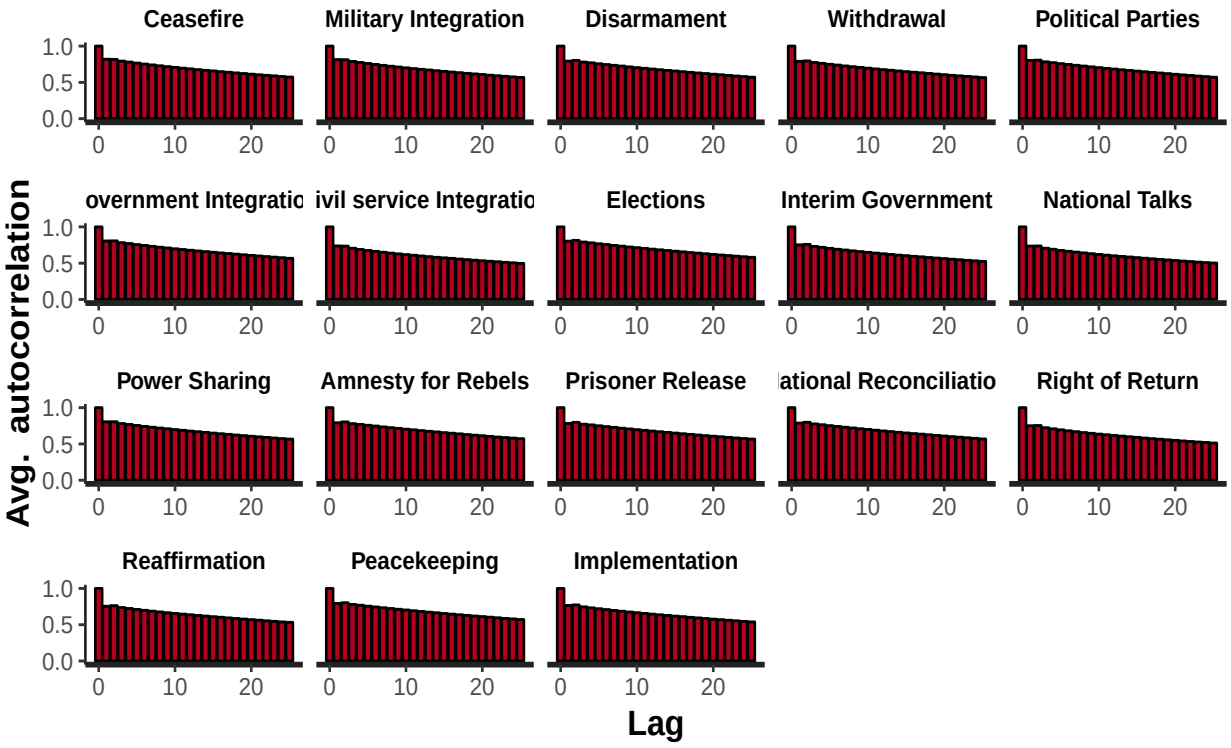
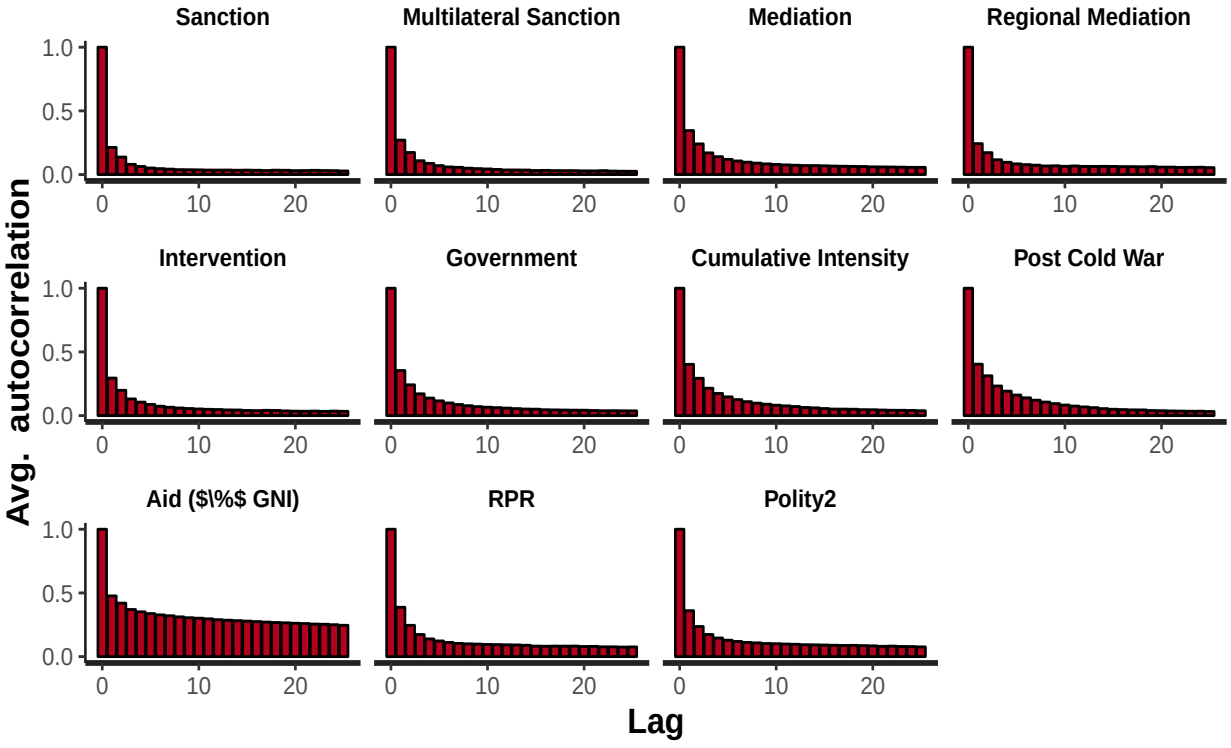


Figure H.4: Traceplot for regression model coefficients in the full probability model. The overlap indicates good mixing between chains.



(a) Discrimination Parameters



(b) Predictor Coefficients

Figure H.5: Autocorrelation of Full Probability Model

I Software Versions

The list below provides version information for the software environment used to generate the results in the paper and SI.

- R version 3.6.0 (2019-04-26), `x86_64-apple-darwin15.6.0`
- 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`
- Random number generation:
- RNG: `Mersenne-Twister`
- Normal: `Inversion`
- Sample: `Rounding`
- Matrix products: default
- BLAS:
  `/Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib`
- LAPACK:
  `/Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: bayesplot 1.7.0, corrplot 0.84, dplyr 0.8.1, forcats 0.4.0, ggplot2 3.1.1, ggrepel 0.8.1, ggridges 0.5.1, knitr 1.23, purrr 0.3.2, readr 1.3.1, reshape2 1.4.3, rstan 2.18.2, StanHeaders 2.18.1, stringr 1.4.0, tibble 2.1.1, tidyr 0.8.3, tidyverse 1.2.1, xtable 1.8-4

- Loaded via a namespace (and not attached): assertthat 0.2.1, backports 1.1.4, broom 0.5.2, callr 3.2.0, cellranger 1.1.0, cli 1.1.0, coda 0.19-2, codetools 0.2-16, colorspace 1.4-1, compiler 3.6.0, crayon 1.3.4, digest 0.6.19, evaluate 0.14, generics 0.0.2, GGally 1.4.0, ggmcmc 1.2, glue 1.3.1, grid 3.6.0, gridExtra 2.3, gtable 0.3.0, haven 2.1.0, hms 0.4.2, httr 1.4.0, inline 0.3.15, jsonlite 1.6, labeling 0.3, lattice 0.20-38, lazyeval 0.2.2, loo 2.1.0, lubridate 1.7.4, magrittr 1.5, matrixStats 0.54.0, modelr 0.1.4, munsell 0.5.0, nlme 3.1-140, parallel 3.6.0, pillar 1.4.1, pkgbuild 1.0.3, pkgconfig 2.0.2, plyr 1.8.4, prettyunits 1.0.2, processx 3.3.1, ps 1.3.0, R6 2.4.0, RColorBrewer 1.1-2, Rcpp 1.0.1, readxl 1.3.1, reshape 0.8.8, rlang 0.3.4, rstudioapi 0.10, rvest 0.3.4, scales 1.0.0, stats4 3.6.0, stringi 1.4.3, texreg 1.36.23, tidyselect 0.2.5, tools 3.6.0, withr 2.1.2, xfun 0.7, xml2 1.2.0

References

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

Gelman, Andrew & Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7(4):457–472.

Geweke, John. 1992. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, April 15-20, 1991*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith. Oxford: Clarendon Press pp. 169–193.

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

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

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

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

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