Guide to Data and Replication Files for *Why Leaders Fight*

If you are interested in using the LEAD dataset, we recommend using the most updated version of the data, WhyLeadersFightLEADDataset_updated.dta.

Dataset Update (12/1/2015): We continued gathering data on leader background experiences after Why Leaders Fight went to press, so we also include replication datasets with updated coding.

- WhyLeadersFightMonadicReplication_updated.dta
- WhyLeadersFightDyadicReplication_updated.dta
- WhyLeadersFightLEADDataset_updated.dta

Replication Overview: This file describes the data and replication files contained in the replication dataverse for *Why Leaders Fight*. The following files are included and replicate the results found in *Why Leaders Fight*. All results were generated using Stata 13.

- WhyLeadersFightMonadicReplication.dta
- WhyLeadersFightMonadicReplication.log
- WhyLeadersFightMonadicReplication_Public.do
- WhyLeadersFightLEADDataset.dta
- Figure5_1.dta
- Figure6_1-6_2.dta
- WhyLeadersFightDyadicReplication.dta
- WhyLeadersFightDyadicReplication.log
- WhyLeadersFightDyadicReplication_Public.do

Erratum: Footnote 56 on page 120 of Why Leaders Fight states that we ran stepwise regression forward and backward with bootstrapped standard errors as a robustness test. Due to convergence issues, we ran stepwise regression on the monadic dataset to generate the reduced dataset. Backward stepwise regression on the dyadic sample, which did converge, produced very similar variable selection (demonstrated in WhyLeadersFightDyadicReplication_Public.do), so we are confident that this did not influence the overall results.

Note on randomization: The randomization used to generate the cross-validated dyadic results is by the leader-year, meaning each leader-year is randomly assigned to the model generation half of the data or the model testing half of the data. Substantively, randomizing on the leader year is preferable. Given the propensity for individual countries like the United States or individual leaders like Adolf Hitler to skew the results, randomizing by the country year is the best way to ensure that each "half" of the data receives a random draw ideal for testing our arguments.

As described in footnote 55 on page 119 in Why Leaders Fight, the results are also broadly consistent when we randomize by the leader or the country (meaning each leader or each country is assigned to the model generation half of the data or the model testing half of the data). This is available in the dyadic replication files.

Note on dataset update: The _updated files also further demonstrates the robustness of the findings in *Why Leaders Fight*. WhyLeadersFightMonadicReplication.dta and

WhyLeadersFightDyadicReplication.dta include some variables set to default values in the absence of confirming or disconfirming information: married, marriedinpower, divorced, and childtotal. Less than 10% of the data was missing for these variables. Well over 90% of the observations of married and marriedinpower were 1. Well over 90% of the divorced observations were 0. We therefore set the values to 1 and 0, respectively, in the absence of available information. We also set childtotal to 3 in the absence of available information, since 3 was the average number of children per leader (once the dataset was reduced down to one observation per leader).

Robustness testing showed that these choices did not influence the results. Moreover, additional research conducted since the book went to press allows us to reverse this coding choice. If you substitute the _updated files in the .do file, the results demonstrate that this decision did not affect the overall results. The last model in WhyLeadersFightMonadicReplication_Public.do replicates Tables 2.3 and 2.3 with the _updated data to illustrate this point.