APPENDIX A
DATA DEVELOPMENT AND CUMULATION FOR INTERNATIONAL RELATIONS: EUGENE

The study of international relations using quantitative analysis relies, in part, on the availability of comprehensive and easily manipulable data sets. To execute large-\(n\) statistical tests of hypotheses, data must be available on the variables of interest, and those data must be manipulated into a suitable format to allow the inclusion of appropriate control variables as well as variables of central theoretical interest. Frequently, however, the process of preparing data sets for analysis is cumbersome, particularly data sets with many cases and with variables that come from a variety of sources. Frequently, control variables are excluded from analysis, not for theoretical or statistical reasons but simply because cumbersome data manipulation tasks preclude optimal test design. The somewhat daunting task of preparing large data sets can have the effect of turning scholars into technicians for substantial periods of time rather than remaining focused on theory development and research design improvements. When data sets are created, errors can creep in, miscodings can occur, and slightly different and poorly documented choices (for instance, how joiners or ongoing disputes are treated in different studies) result in nonreplicable data sets (sometimes by the data set creators themselves). For many scholars the barrier to entry into, and becoming competent in, the realm of quantitative research is sufficiently high to preclude any sophisticated analysis at all.

These issues are particularly serious when assembling data sets for comparative theory testing. Our data sets are dyadic, large, and contain many variables, and we made a large number of specific research design choices to fit our theoretical approach. But our data set is clearly not the last data set that will ever be required in international relations; to the contrary, we believe that changes, expansions, and extensions are
essential to furthering the project we began here. While we believe our choices are the most appropriate for our task, others may disagree. Some may wish to use different rules for case inclusion and exclusion. Some may find other theories that we have not included particularly important and may wish to include them. Others may wish to operationalize variables differently for the theories we do analyze. Finally, new theories of conflict will certainly be developed in the future that scholars will wish to test against existing explanations for conflict. Making selected modifications while maintaining most of what we did is a daunting task.

To allow replication and easy extensions of our work, and to lower the barriers to entry in quantitative analysis, we have developed software designed to eliminate many of the difficulties commonly involved in constructing large international relations data sets, particularly dyadic data sets. Our software, titled EUGene (the Expected Utility Generation and Data Management Program), also generates the data necessary to test the international interaction game version of the expected utility theory of war developed in Bueno de Mesquita and Lalman 1992. EUGene is a stand-alone Microsoft Windows–based program for the construction of annual international relations data sets and is designed to make building such data sets simple. It accomplishes this by automating a variety of tasks necessary to integrate several data building blocks commonly used in tests of international relations theories. By reducing the time necessary to carry out routine data set construction tasks, EUGene allows users to proceed more rapidly to the analysis stage and allows scholars to spend more time on theory development and on asking new research questions than on data management. It also facilitates replication by providing a single program for data set creation that will produce the same results for all users, eliminating the problem of hidden or forgotten steps typically encountered when attempting replication.

Usage Overview

Updates, details, and the material needed to replicate our results are available free of charge from the Inter-University Consortium on Political and Social Research (www.icpsr.org, study number 1290).

Users will normally use one of three main choices when they enter the program. Most often, users will choose to construct an output data set containing any or all of the variables discussed previously by using the “Create Data Set” menu. A tabbed window provides the option for users to set the unit of analysis, population of cases, variables, and output for-
mat for their output data. EUGene also forces choices on a variety of critical but often unstated assumptions about the construction of key dependent variables and the inclusion of problematic cases that go into the construction of international relations data sets (how to treat ongoing disputes, for example). The program then assembles an output data set according to these user specifications, handling necessary merges between different input data sets and creating command files to automate reading the data into other statistical programs. Second, users may recompute expected utility data (or various components of it) under the “Recompute” menu. Usually users do not need to do this, as we have already precalculated data where appropriate, and all variables available for selection within the program are ready for output. The recalculation options exist for users who wish to create new data sets to examine the sensitivity of the results to various assumptions; submenus under “Recompute” allow users to specify some of those conditions, such as what distance discounting method to use. Recalculating these data, especially recalculating risk attitude data, is quite time consuming. As a third possibility, users may decide to upload or download new add-on data sets for EUGene under the “User Data” menu. Downloading new data makes variables from other users’ data sets available in the program. Users may also upload their own data sets to www.eugenesoftware.org for eventual inclusion as possible downloads by others.

Unit of Analysis

The first choice made by users when creating a new data set in EUGene is the unit of analysis. Users choose to create data sets with the country-year, directed dyad-year, nondirected dyad-year, directed dispute-dyad, and directed dispute-dyad-year as the unit of analysis. By selecting these units of analysis, users can examine monadic time-series (by creating a country-year data set), examine dispute initiation from a condition of peace or examine the duration of peace (by creating a directed or nondirected dyad-year data set), examine the escalation of disputes (by creating a directed dispute-dyad data set), or examine the evolution and duration of disputes over time (by creating a directed dispute-dyad-year data set). Clearly, EUGene is particularly useful when creating data sets with the dyad-year as the unit of analysis. Scholars have increasingly come to use data sets based on the dyad-year to conduct quantitative analyses. This is because dyadic interaction lies at the heart of strategic international behavior and because it is possible to combine explanations from multiple levels of analysis in one quantitative study, as we
have done here. Most scholars rely on annual data both because data is widely available at this level of temporal aggregation and because the year represents a natural political break due to budget cycles, electoral cycles, and the presence of winter or a rainy season that in many areas hampers military action.¹

Variables

Users specify the variables that are to be included in the output data set by clicking on a set of check boxes. The program as currently distributed allows users to choose from a set of over sixty variables from several of the most important international relations data sets. Variables that can currently be selected for inclusion include Polity III democracy scores and ancillary components (Jaggers and Gurr 1995), COW project capability data (Singer, Bremer, and Stuckey 1972), data on interstate distances, alliance data, tau_b scores, $S$ scores (Signorino and Ritter 1999), risk attitude data, contiguity data, region, peace years (Beck, Katz, and Tucker 1998), expected utility values and international interaction game equilibrium predictions (Bueno de Mesquita and Lalman 1992), COW MID data (Jones, Bremer, and Singer 1996), and the Maoz dyadic implementation of them (Maoz 1999). Users may also download additional data sets that have been submitted to the EUGene Web site by other users of the program. They may also upload new data to our Web site for inclusion in EUGene. When such user data sets are placed in the appropriate program directory, EUGene automatically makes the variables that they contain available for selection.

Population of Cases

The scope of output data from EUGene may be set as either all dyad-years or some specified subset of countries and years. Users can specify a particular range of years for output (e.g., 1945–92 or 1816–1914) and can select from commonly used subsets of countries (e.g., all dyads, politically relevant dyads, major power dyads, contiguous states, or a user-selected list such as rivals). Alternatively, users may generate all dyad-years and include variables in the output (such as a “politically relevant” dummy marker) to allow selection at a later time. The creation of various case subsets allows users to conduct contingency analyses and to explore the sensitivity of their results to factors like era or region.
Case Inclusion Criteria and Assumptions

EUGene forces users to make specific choices concerning the inclusion or exclusion of potentially problematic data points related to dependent variable codings and case censoring. EUGene forces choices on three specific issues. The first concerns years with ongoing militarized disputes. EUGene allows users to either drop or include dyad-years where the countries begin the year with an ongoing dispute (users may want to drop such cases if they believe that a new initiation would be censored by the ongoing dispute).\(^2\)

The second issue concerns the treatment of dyads where a state joins into a dispute that is already in progress. Should joiners be included for analysis in the same way as dispute initiators? We have argued that the information conditions faced by joiners into disputes is fundamentally different than the conditions facing the initial participants and omit joiners from analysis. Others may disagree. EUGene allows users to include or to drop such cases by selecting a check box.

The final issue concerns “target versus initiator” directed dyads. When one state initiates a dispute, it does so against a target state, creating a designated initiator A and target B. But when A initiates versus B, it is less than clear how to include the directed dyad B versus A, because true behavior in the B versus A dyad may be censored. EUGene gives users the option to include such target versus initiator dyads in the data sets it creates, to drop them, or to include them only if there is a subsequent initiation by B versus A.

Merging and Data Conversion

One difficulty with building data sets that combine variables is that input data sets frequently come with different units of analysis and in different formats, requiring conversion at a fundamental level in addition to simply merging. For example, many key IR data sets have the country-year as the unit of analysis (e.g., the COW national capability data, Gurr Polity data, or data on national risk attitude). Other data sets (or data constructions) have the dyad as the unit of analysis, such as distance data, the COW contiguity data set, or data on expected utility. Still other data sets are distributed in a hybrid form, such as the COW MID data set, which is dyadic and annual in its underlying form but comes distributed as three separate files that must be merged together. EUGene carries out necessary conversions among the formats,
file structures, and differing units of analysis of these data sets automatically as part of the merging process.

**Dispute Data Conversion to Dyads**

The COW MID data set (Jones, Bremer, and Singer 1996) is commonly used to create most dependent variables in recent quantitative international relations studies. However, prior to 2003 the COW project has not distributed this data in a dyadic form. Converting the data to a dyadic format involves checking states for their involvement as originators or joiners, identifying states as initiators or targets (the initiator is the side of the MID that first crosses the militarized threshold, side A in the MID data), and pairing into dyads. Appropriately pairing states as “real” dyads rather than simply participants on opposite sides of a multilateral dispute involves checking to be sure that the states are actually involved on opposite sides at the same time (some simple conversions do not perform this check). The procedure to make appropriate pairings from the COW MID version 2.1 data set (Correlates of War 1999) involves reading data from three files, two of which have the dispute as the unit of analysis (one record per dispute) and one of which has the country involvement as the unit of analysis (one record per state involved in the dispute per dispute). Country-dispute-level information in the data set must be matched to the dispute-level data, and multiple country-dispute-level records must be matched to each other to obtain dyadic pairings. EUGene does this automatically.

Zeev Maoz has recently developed a dyadic version of the MID data (Maoz 1999) that is designed to correct problems involved in converting the original MID data into dyadic format. While Maoz begins by converting the original MID data sets into a dyadic format, much as EUGene does, he then conducts a variety of additional checks intended to verify that actual disputatious interaction occurred between the members of the dyad. Maoz’s data are available as an alternative version of the dispute variables within EUGene.

**Dependent Variable Coding**

Scholars have made a number of arguments about the appropriate unit of analysis in international relations studies and about coding the origins of militarized disputes as a dependent variable in those analyses. In a directed dyad setting, the initiation of a dispute is the appropriate
coding of dispute. The COW MID operationalization of initiator in practice is that the initiator is the “first mover,” that is, the state who first crosses the MID threshold and makes the first threat or actual use or movement of forces. This definition of “initiator” gains clarity in terms of the temporal ordering of actions while losing any attempt to get at intent. An alternative image of the initiator that comes to mind is the predatory state that decides to engage in conflict against a state that wants to remain at peace. This may or may not be the state that moves first. EUGene allows users to specify an alternative coding for dispute initiators (which also comes from the MID data set) as those who are “revisionist” states.

A second issue has to do with whether the directed initiation of militarized disputes can be measured in a meaningful fashion (see the discussion in chapter 2). If users select nondirected dyads as the unit of analysis, the occurrence of MIDs is automatically measured and output as dispute “onset” rather than dispute “initiation” as it is in directed dyad data sets.

A final issue of dependent variable coding lies in the treatment of disputes that continue for more than one year. We have argued that only the first year of a new MID should be coded as a dispute initiation and that subsequent years of multiyear disputes should be dropped. Others who believe that our models should do equally well at predicting the continuation as well as the initiation of MIDs argue that we should code ongoing dispute years as a “1” as well as just the first year. EUGene allows users to specify either that only the first year of a MID is coded as a dispute initiation or that all years of a continuing MID are coded as dispute initiations.

Output Format and Use of Other Software

EUGene is not an analysis program but rather a data management utility. As a result, the merged data created by EUGene must be read into and analyzed by other statistical software. Users may then conduct analysis immediately or more likely will wish to compute additional variables on the basis of the data created by EUGene. EUGene’s output files are created in a uniform format that can be read easily into any statistical analysis software. Data may be tab-delimited (tab characters are placed between values in the data file), space delimited, or comma delimited. EUGene also creates the command files necessary to import the data into SPSS, STATA, or LIMDEP; the command files both read the raw variables and set missing values appropriately. After creating a data
set, users can then have the data up and running in these statistics programs in a matter of minutes.

**Expected Utility Data**

Users of EUGene may take advantage of the expected utility data sets and variables that come with the software. In its life as a data generation program, EUGene generates expected utility data for all dyads and years for which raw input data are available, following the methods developed in Bueno de Mesquita and Lalman (1992) and starting from first principles. Any interested user can view the complete code for these calculations in the form of EUGene’s source code. EUGene actually has several intended uses vis-à-vis expected utility data. Most obviously, the availability of complete expected utility and IIG equilibrium predictions allows us to test arguments about the IIG and the expected utility theory of war in a much broader setting. A key goal was to implement the methods of previous expected utility calculations, which Bueno de Mesquita and others published across a number of books and articles, in a single, easily accessible package. Expected utility data is generated from first principles using the most recent updates to the COW alliance, national capabilities, and MID data sets to ensure that researchers have the most accurate expected utility estimates possible and to ensure transparency and further replication. By making data on all dyads available, EUGene also allows us to test previously unexamined arguments about expected utility. For example, with data generated by EUGene, it is possible to examine systematic differences across regions or time periods or to look for systematic variation in risk-taking propensities between countries.

**Technical Information**

EUGene runs under Microsoft Windows 95, 98, NT (version 4.0 or higher), 2000, and XP on IBM-compatible PCs with at least 16 MB of memory. The program has a standard Windows program interface. EUGene is written in Borland, Inc.’s Delphi programming environment, an object-oriented Rapid Application Development package designed to create Windows programs, relying on a Pascal base. Version 2 of the program consists of approximately 36,000 lines of computer code split into eighty-five units and Windows forms. The program is copyrighted but is available as a free download from www.eugenesoftware.org.
downloadable program includes the main program executable file
(about 1.4 MB), complete (eighty-page) documentation of the program,
EUgene’s source code, and the complete expected utility data described
here. In addition to providing instructions for how to use the program,
the program documentation further details the computations involved
in making expected utility calculations, lists the variables available for
output, and outlines key algorithms used in the program.