International Interactions, 30:109–141, 2004
Copyright © Taylor & Francis, Inc.
ISSN: 0305-0629
DOI: 10.1080/03050620490462603

MEASURING THE INTENSITY OF INTRANATIONAL POLITICAL EVENTS DATA: TWO INTERVAL-LIKE SCALES

STEPHEN M. SHELLMAN

Department of Government, College of William & Mary, Williamsburg, Virginia, USA

King, Keohane, and Verba (1994, pp. 27–28) argue that we should "maximize the validity of our measurement," "ensure that data-collection methods are reliable," and make all data and analysis replicable. In an effort to improve the measurement of the events data collected by the Intranational Political Interactions (IPI) project, this extension of the project produces two new valid and reliable interval-like scales. Following Azar (1982), Goldstein (1992), and Moore and Lindstrom (1996), I produce interval-like scales of cooperative and hostile political actions based on a group of experts' judgements. The collective scaling procedure produces data suitable for use in OLS regression models as well as a standardized interval-like scale that more accurately represents the true scores of event types. The paper discusses the procedures taken to derive the new measures, proceeds to argue why the new measures are improvements over existing measures, and reports the findings of statistical analytic comparisons. The statistical comparisons demonstrate that the new scales make a difference in various statistical models using different temporal units of aggregation.

KEY WORDS: magnitude scaling, measurement error, conflict, cooperation, temporal aggregation, event data

INTRODUCTION

Scholars who study social actors' behavior have an interest in theorizing about causal processes and testing those theoretical propositions using statistical inference.

Received for publication 1 April 2002.

I wish to thank the expert judges who collectively weighted the events. In addition, I would like to thank the anonymous reviewers, Bill Berry, Ashley Leeds, Andrew Long, Sara Mitchell, and Will Moore for useful comments and suggestions. A previous version of this paper was presented at the 2000 Peace Science Society (International) meeting, Yale University, New Haven, Connecticut, October 27–29.

Address correspondence to Stephen M. Shellman, Department of Government, College of William & Mary, Williamsburg, VA 23188-8795, USA.

To draw valid inferences about social behavior from our statistical analyses we need valid and reliable data. In short, measurement matters. Most would agree that the results obtained from econometric models depend, in part, on the measurement of the variables. In cases where one measure is more valid and reliable than another, scholars should employ the more valid and reliable measure. King, Keohane, and Verba (1994, pp. 27–28) argue that we should "maximize the validity of our measurement," "ensure that data-collection methods are reliable," and make all data and analysis replicable.

In this study, I pursue King, Keohane, and Verba's argument and concentrate on improving the quality of a domestic conflict-cooperation events data set known as the Intranational Political Interactions (IPI) project. Event data—"day-by-day coded accounts of who did what to whom as reported in the open press"—offer the most detailed record of interactions between and among actors (Goldstein, 1992, p. 369). To utilize event data in statistical models, one must convert individual events to time series and aggregate them in a way that requires some method of combining different event types into a "single theoretically meaningful measure (in one or more dimensions)" of the relationships among actors (Goldstein, 1992, p. 370). Most event data projects convert events into a measure of conflict or cooperation (e.g., Cooperation and Peace Data Bank—COPDAB, World Events Interaction Survey—WEIS, Integrated Data for Events Analysis—IDEA, Protocol for the Assessment of Nonviolent Direct Action—PANDA). COPDAB includes both a cooperation scale and conflict scale that were developed by judges (Azar, 1982) and they may be "joined" to form a single dimension of conflict-cooperation (Sloan, 1975, pp. 31–32; Goldstein, 1992, p. 370). In the same way, the WEIS events were scaled by both Vincent (1979) and Goldstein (1992) to form a conflict-cooperation dimension. I perform a similar exercise using the event categories of the Intranational Political Interactions (IPI) project.

IPI is an events data project designed to study the behavior of political actors within societies.[1] While the project proves useful in the study of such behavior (Moore and Davis, 1998; Lee, 2001), the ordinal IPI scales impose on the data a constant, interval, linear relationship between the events and their assigned values.[2] This extension to the IPI project takes issue with the ordinal measures and seeks to accomplish three particular objectives.

First, the extension converts an ordinal data set into an interval-like data set so that the data are appropriate in ordinary least squares (OLS) regression models. When we use OLS regression models to test the implications of our theories, the method assumes that the variables are measured on an interval scale (Berry and Feldman, 1985, p. 10). Thus, I take steps to transform the existing IPI *ordered* data scheme into an *interval-like* scheme.

Second, the extension produces a *common* domestic conflict-cooperation interval-like scale to make analyses replicable.[3] In the past scholars assigned their own intensity weights to the ordered categories producing different weighting schemes. I take steps to resolve the inconsistencies among the different scales.

Finally, the extension maximizes the validity and reliability of the IPI event data set. Should we impose the rather strong assumption that the intervals between each event category are the same? Rather than impose such a strong assumption, I choose to survey a set of scholars who work in the field to determine whether this linear assumption has an empirical basis. This research follows the work of Azar and Sloan (1975), Goldstein (1992), and Moore and Lindstrom (1996) in that it asks political violence experts to collectively order and assign intensity weights to the event categories of the IPI project. The panel make-up and the reported reliability checks suggest that the resulting data are valid and reliable.

The survey produces two common *interval-like* scales of domestic cooperative and hostile political actions based on a group of experts' judgements. I show how the expert scale differs from the ordinal scale using graphic representations and then proceed to illustrate the differences between statistical results obtained using the ordinal IPI scale and results obtained using the new interval-like scale. Using two general statistical models (i.e., Richarson Arms Race type models and VAR models) and estimation methods (two-stage least squares or 2SLS and OLS) encountered in the literature, I estimate one set of parameters using the ordinal measures and another set of parameters using the interval measures. In addition, I control for the problems associated with case bias and temporal aggregation by estimating each model using data from two IPI cases (Chile and Venezuela) which are aggregated by three different temporal units (days, months, and turns).

The paper begins with a note on the purpose and significance of the IPI project and an explanation for interval-like scale construction. Next, I describe the procedures taken to derive the new measures and report the new IPI weights along with their reliability coefficients. Finally, I discuss similarities and differences between the old and new scales and show that one obtains different parameter estimates and significance levels when comparing statistical results across the two scales.

THE INTRANATIONAL POLITICAL INTERACTIONS PROJECT

While event data focussing on international actors' behavior are fairly common (WEIS and COPDAB), relatively few data collection efforts code domestic actors' behavior. One exception is the IPI project.[4] The IPI project is designed to measure political conflict and cooperation within societies by coding political events reported in the news (Davis, Leeds, and Moore 1998). Many scholars argue that all actors' behavior while engaged in political interactions can be placed somewhere on a hostility-cooperation (H-C) continuum (Goldstein and Freeman, 1990, 1991; Moore, 1995, 1998, 2000). IPI adopts this view and codes news events into two scales, each one consisting of ten ordered categories, based on the intensity and severity of cooperative or hostile actions. The IPI project collects event data over the years 1979–1992 in 14 middle power countries.[5]

DRAWBACKS TO IPI

The IPI data prove useful in the study of dissident responses to state repression (Lee 2001) and in the exploration of the domestic-international conflict nexus (Moore and Davis 1998). However, the major drawback of the existing IPI scales is that the measures are ordinal. One assumption often invoked in least squares estimation is that variables are measured at the interval level (Manheim and Rich, 1995, p. 330; Berry and Feldman, 1985, p. 10).[6] The Pearson correlation coefficient is attenuated when computed with non-interval level data, and since it is used in numerous multivariable methods, such as multivariate regression, the concern is magnified (Schumacker, et. al., 2001, p. 2). Both Wilson (1971) and Marks (1974) argue that the use of ordinal-level data in interval-level statistical techniques is not legitimate and is not empirically defendable.[7]

Therefore, to properly use the IPI data in OLS regression models, the ordinal measures must be transformed in some manner to interval-like data. In previous research, both Lee (2001) and Moore and Davis (1998) take steps to meet the interval-level measurement assumption in their regression analyses by assigning their own interval weights to the ordered IPI categories. Moore and Davis (1998) converted the IPI data to a scale Moore and Lindstom (1996) created for the *Violent Intranational Conflict Data Project* (VICDP)[8], while Lee (2001) assigned his own subjective weights. These different conversion processes lead to two different scales. One purpose of constructing the new measures is to provide a common, valid, and reliable interval-like scale that can be used in statistical analyses that assume interval-level measures.

A second assumption invoked in least squares estimation is that variables are measured perfectly, without error (Manheim and Rich, 1995, p. 330; Berry and Feldman, 1985, p. 10). I show below that the IPI categories artificially over and under represent hostile and cooperative events. That is, the events are measured with error. The result is inefficient and biased parameter estimates of domestic actors' behavior. I discuss this issue in more detail in the next section.

MEASURING VIOLENCE WITH ERROR AND ITS STATISTICAL IMPACT

According to Berry and Feldman (1985, p. 26), "Measurement error is one of the most important methodological problems facing the social sciences, and it can have a major impact on the estimation of regression coefficients in otherwise well-specified models." When we run statistical models, we assume that our variables are measured perfectly, without error (Berry, 1993, p. 12). However, in many cases this is a strong assumption. In the real world, data are often contaminated with error, or researchers use indicators that are mere proxies for rich theoretical concepts. Thus, many data we collect and employ in statistical models are subject to error.

If variables are measured with error, adverse consequences may arise. According to Bollen (1989, p. 166–167), in the multiple regression context, "if only one explanatory variable is subject to error, its regression coefficient is attenuated …if more than one variable is measured with error … coefficients from equations that ignore error in variance can be higher, lower, or even the same as the true coefficients." Below, I specify a formal measurement model that helps communicate how error contaminates quantitative measures.

In most instances we use an indicator to represent a concept. Each indicator is a function of its true score and some error.[9] For example, suppose the true score (T) of the concept "state behavior" is measured by an indicator (E) (observed event score), where the event score for any state action, j, directed toward a target is a function of the true score and an error term, v:

$$E = f(T_j, v_j) \tag{1}$$

As is well known, the error, v, may be of two types, random and nonrandom. The type of error determines the consequences in regression analyses. For example, random measurement error (RME) in an independent variable of a regression model biases parameter estimates such that the amount of bias is a function of the magnitude of the measurement error and the co-linearity among the independent variables (Berry, 1993, p. 51; Berry and Feldman, 1985, p. 28–31). "Non-random measurement error (NRME) will always lead to bias in OLS estimators" (Berry, 1993, p. 51). Non-random error may be linear or nonlinear, where linear refers to the indicator over- or underestimating the true score by a constant percentage, absolute amount, or both, and nonlinear refers to error as a function of the variable being measured. This relationship may take on various functional forms. However, only if we know the true score can we really assess the bias and inefficiency that measurement error causes. Our best strategies for dealing with measurement error are to theoretically clarify each concept, to produce highly valid and reliable measures for our concepts, and to employ such measures in our analyses.

The following section describes the procedures taken to improve the measurement of the ordinal IPI scales. I begin with conceptualization issues and move on to discuss how I selected individuals to perform the weighting procedure. Next, I describe each phase of the scaling procedures. Finally, I discuss how I assess the reliability of the scales.

RESEARCH DESIGN

Conceptualization of the Hostility-Cooperation Continuum

The IPI project seeks to measure political actors' behavior directed towards one another and relies on descriptions of political events to code such behavior. The IPI project conceptualizes and treats behavior in terms of a least-to-most dimension that ranges from highly hostile actions to highly cooperative ones. The Leeds, Davis, and Moore (LDM) scale designates 20 ordered categories along this hostility-cooperation (H-C) continuum. In particular, it consists of a 10-point ordered scale of hostility fused with a 10-point ordered scale of cooperation. The hostility scale ranges from least (10) to most (100) hostile political actions, while the cooperative scale ranges from least (10) to most (100) cooperative political actions (See Tables 1 and 2). The ordered structure to the event categories is important for studying *levels* of and/or *change* in actors' behavior. A simple count of events deemed cooperative and/or hostile is useful in studying the frequency of cooperative and/or hostile behavior, but dummy variables or event counts make it more difficult to draw inferences concerning actors' levels of behavior during contentious and/or cooperative interactions with one another.[10]

While the idea that behavior can be placed somewhere on a H-C continuum is sound, the idea that the categories on this scale are the same number of units apart is not empirically obvious. An example should help to illustrate. The LDM scale assumes that the distance (10 units) on the hostility continuum (ranging from 0 to 100)

Table 1
LDM Hostility Scale

Event (Value)	Event Description
1 (10)	Mildly negative statements (verbal or printed) about other parties, their representatives, proposals, or activities. No action threatened or implied.
2 (20)	*Strongly* negative statements representing implicit or explicit threats. Rejections of proposals.
3 (30)	Nonviolent protests, demonstrations, or strikes with political intent. Minor restrictions on political and economic participation. Legal actions protesting leadership or policies.
4 (40)	Minor political violence and more significant restrictions on political and economic participation. Riots, violent demonstrations, property damage, no deaths. National strikes.
5 (50)	General restrictions on political and economic participation, and political violence. Imposing regional curfews. Legal actions ending tenure of ruling group.
6 (60)	Illegal attempts at ending tenure of ruling group or extra legal violent activities. Riot or police violence in response to protest or riot involving 11 to 99 deaths.
7 (70)	Extensive political violence. Major rioting or government violence in response to protest or riot involving hundreds of deaths.
8 (80)	Changing the structure of government or very high levels of political violence. Ending normal governmental policy/decision making process. Executive suspension of the legislature.
9 (90)	Societal upheaval. Rebels setting up rival government. Declaration of martial law. Violent coup that alters the structure of government.
10 (100)	Civil war, genocide/politicide, major battle, violent coup followed by purge.

between a mild negative statement (10 on LDM scale) and a strong negative statement (20) is the same as the distance (10 units) between large-scale rioting (70) and changing the structure of the government (80). In short, all categories are 10 units apart on a scale ranging from 0 to 100. Empirically, most conflict–cooperation scales of actor's behavior, such as Goldstein's (1992) WEIS scale and Azar's (1980) COPDAB scale are nonlinear (e.g., the intervals between event types are not constant).

Both Lodge (1982) and Shinn (1974) argue that categorical judgments do not routinely reflect intervals. That is, the magnitudes are not properly measured using constant interval distances. Rarely do continuous measures accurately reflect ordinal measures of the same concept. Similarly, I argue that the constant distance between

Table 2
LDM Cooperation Scale

Event (Value)	Event Description
1 (10)	Mildly positive statements, verbal or printed, about other parties, their proposals, or activities. Statements of support
2 (20)	Strongly positive statements about other parties, their representatives, proposals, or activities. Implied or literal promises of action. Promises directed at particular actors.
3 (30)	Minor cooperative actions. Relaxing individual restrictions on free speech/ free press/individual action. Agreements between groups that have not consistently cooperated in the past. Helping another group to mobilize resources/gain political advantage.
4 (40)	Agreements to attempt to settle protracted conflict or relaxing minor restrictions. Relaxing travel and movement restrictions; ending local curfew.
5 (50)	Relaxing government sanctions or actions designed to mitigate protracted conflict. Releasing hostages. Allowing regional elections to take place.
6 (60)	Reforms; relaxing major restrictions; truces. Allowing opposition parties to take power following elections. Relaxation of major repressive activities. Implementing policy reform.
7 (70)	Substantial agreements. Ending nationwide state of emergency. Convening a commission to write new constitution (multiple parties/groups involved). Peace treaties.
8 (80)	Conflict termination: Hold elections under old constitution in which all parties to the conflict participate.
9 (90)	Conflict settlement: Implementing a new constitution that guarantees political and civil rights to all participants in the conflict or one that grants autonomy to specific groups.
10 (100)	Conflict resolution. The internal war is terminated because the underlying conflict is resolved such that each party’s needs are guaranteed.

categories in the LDM scale introduces measurement error and that one must conceptualize a H-C continuum in which categories are not the same distance apart from one another. In an effort to improve the accuracy of the IPI measures, this project surveys experts to weight IPI event categories in an interval-like fashion. The result is a H-C continuum that accounts for differing distances between categories. Next, I introduce how I selected the panel to place event types along this H-C continuum.

Panel Selection and Issues of Validity

Why does the makeup of the panel matter? The panel must be capable of distinguishing the relative intensities of the particular variables (Zaninovich, 1963, p. 56). In this case, the relative variables are *hostility* and *cooperation*. Specifically, the panel must distinguish among domestic political actors' levels of cooperative behavior and hostile behavior. Rather than take a representative sample of all types of individuals to make these distinctions, I choose to sample from a group of well-known internal conflict scholars. Drawing on knowledge of internal political conflict, an expert should be able to make the distinctions better than an average person. Because I am interested in how experts perceive the intensity of domestic political acts, I drew a sample from scholars around the world who specialize in the study of internal conflict.[11]

Although I wish to access knowledge from experts, I want to represent different types of backgrounds on the panel. Campbell and Fiske (1959) define validity as the agreement among maximally different methods for measuring a concept.[12] This view of validity suggests that one should survey political violence experts from multiple disciplines and career points to perform the task. Different people's perspectives from different backgrounds are used to measure the concept and are viewed as multiple methods.

Specifically, the panel consisted of political scientists and sociologists who have published in the area of political violence and who are regarded as experts in the field.[13] I concentrated on choosing research faculty who had extensive knowledge on social movements, dissent and repression, revolutions, and ethnic and religious conflict within countries.[14] In addition, many, but not all, scholars had some attachment to working with quantitative data (event data in many of those cases) and statistical models. Finally, these experts are at various points in their careers such as full professors, associate professors, assistant professors, and even one advanced doctoral student. I solicited 25 "experts," of which seventeen (68% response rate) completed the scaling and weighting procedure and returned the results to me.[15] Comparing the characteristics of those sampled (25) and those who returned the results (17), the respondents are representative of the sample population.[16] Having discussed the panel, I move on to discuss the stimuli that the judges measured and the technique employed.

The Stimuli (Events)

Largely based on Q-Sort, I devise a technique that is useful in comparing and measuring intensities of behavior across a series of stimuli. The stimuli are unit-statements or event-statements that describe domestic political acts. The statement is said to "stimulate" a reaction by the scientist to evaluate where this political act falls on the scale—in this case ranging from 1 to 10 for hostility and 1 to 10 for cooperation. The unit-statements should be stated in such a way that all references to "particular countries, decision-makers, and geographic locales be eliminated" (Zaninovich, 1963, p. 58). An example of such a unit statement is the following: "government X ends local curfew in region Y." This in effect masks the content so that judges' prejudices are reduced in the scaling operation. For example, a scholar who identifies with a particular group or regime due to his or her ethnic, religious, or cultural background, may introduce prejudice into the process. By not specifying a group or regime, there is less prejudice introduced into the scaling procedure.

Table 3
Events Weighted by Judges

Code	Cooperative Events
CE	Group X praises Institution Y for policy implementation.
CJ	Government X promises Group Y that it will make concessions to end the dispute.
CB	Group X signs pact with Government Y. Pact promises to end a minor dispute between X and Y.
CG	Government X ends local curfew in Region Y.
CI	Government X agrees to allow regional elections to be held in Region Y.
CC	Government X allows opposition Group Y to take power following the election.
CF	Government X ends the nationwide state of emergency. Population Y is effected.
CH	Government X holds first national elections in 14 years (under old constitution). Population Y is allowed to participate.
CD	Government Y implements new constitution that guarantees political and civil rights to Group X.
CA	Internal war between Group X and Government Y is terminated due to a resolution. both X's and Y' s needs are guaranteed.
	Hostile Events
HD	Group X criticizes Government Y's new policy.
HA	Government X rejects opposition Group Y's proposals for reform.
HG	Group X demonstrates nonviolently against Government Y.
HH	Group Y riots in response to Government X's policies (property damage, 0 deaths).
HC	Government X imposes regional curfew on Region Y.
HF	Government X fires into crowd of Group X protesters (40 deaths).
HJ	Election violence in region Z between Government X and Group Y (100 deaths).
HI	Government X suspends the national constitution. Population Y is effected.
HB	Group X violently topples President Y's government; X installs itself in power.
HE	Government X executes hundreds of members of Group Y.

To mask the content, countries, specific political groups, leaders, etc. are substituted with symbols. For example, the statement "the military has taken over for President Shugari in Nigeria" is replaced with "group X has taken over for President Y in state Z." A list of the events that were presented to the judges is displayed in Table 3.

The Procedures

Each expert received a "judging packet" which included a letter asking for their expertise and participation as a judge, instructions for scaling, an envelope containing 10 slips of paper, each representing a hostile event, a second envelope containing another 10 slips of paper, each representing a cooperative event, two separate worksheets which served as categorical "game boards" for laying out the events from least to most in order to scale, and a results form to record the order of events and corresponding weight assignments.[17] Each judge is asked to scale 10 generic cooperative events from least to most cooperative and 10 generic hostile events from least hostile to most hostile. Then, they are asked to assign to each event a weight ranging from 1 to 100. The procedure takes place in three phases for both the hostile and the cooperative events.

Phase 1 In the first phase, the scaling judge is asked to open the envelope marked hostile and freely place the events in the slots on the "Weighting-Scale" in a *least- to-most* rank-order. The judges are to assume that the events are taking place between a nation state's government and a group that challenges the government's authority to rule.[18] Additionally, the judges are told to concentrate on the "level of behavior" or "action" that is exemplified in the event and not on the group who is performing the action. Thus, the focus is on the behavior and not the actors.

I ask each judge to place one statement into each category. In other words, these statements are forced into a ten-category differentiation. Accordingly, a rank-order of classes of statements is derived. This prevents the judge from classifying two or more events in the same category. It allows for a neat fit of the statements into categories and results in a more consistent scale.[19] The forced correspondence of one statement into one category serves both to impose discipline upon the judges and to provide a common base for comparing the intensity of a set of event types. Moreover, the procedure takes less time and imposes less of a burden on each judge.

However, it is important that each statement represents the IPI category well from which it is derived. For example, if the category heading is "nonviolent protests, demonstrations, or strikes," the unit-statement should contain the words "nonviolent protest," "demonstration," or "strike." I carefully chose statements that were well representative of the categories they were delineating. Because each event represents its corresponding category (what it intends to represent), the events are valid indicators of each category.

Phase 2 In the second phase, the judge is asked to inspect his or her initial scale, make comparisons, and shift the statements around until he or she is satisfied with the rank-order of the statements. The judgment is to be made in terms of the relation of one statement to another.

Phase 3 At the end of phase 2, the events are in rank-order, thus producing an ordinal scale. However, the goal of this project is to produce interval-like scales. Therefore, in the third phase, each judge is asked to weight each of the events. In other words, the judges assign a value, ranging from 1 to 100, where 1 is least hostile and 100 is most hostile, to each of the events. The judges are again asked to compare events to one another and check the intervals among event types (i.e., if one judge assigns a 22 to one event and a 44 to another, is the second event type twice as hostile (or cooperative) as the first). Finally, the judges are asked to repeat the procedure for the cooperative events, compare the two lists to one another, check the interval distances across lists, make any necessary changes, and record their results on the "results form," which was then returned to me.

Reliability

The reliability of the scales is established by the degree of agreement among the sample of judges for the universe of statements that were considered (Zaninovich, 1963, p. 62). The degree of agreement is determined by calculating both an inter-judge correlation and a composite reliability score. Since I am asking the judges to assign weights to events, creating an interval-like scale, I focus on the reliability of the interval-like scale as opposed to the ordinal scale. Since the weighted data are interval-like measurements, I calculate the inter-judge correlation using the Pearson product moment correlation coefficient.[20] White and Saltz (1974) suggest that "there is no reason why the techniques of computing reproducibility cannot be reversed to yield coefficients about the homogeneity of subjects instead of test items" (p. 193). In effect, this is exactly what I do; I determine the homogeneity of the sample of judges using the Pearson formula.

Since the mean of the entire sample of judgments (weighted values) on each of the 20 statements is used for statistical analyses, the *composite reliability* regarding the consensus of all judges is the most important indicator of the scale's reliability (Zaninovich 1963, p. 64). In other words, if we were to draw another sample and carry out the procedures, would this second draw correlate with the one sampled in this study (Block, 1961, p. 37). I use Cronbach's alpha to calculate the composite reliability. Block (p. 38) suggests that a composite reliability of +0.80 is adequate for data analysis.

THE NEW SCALES

The procedure produced two new interval-like scales. First, I produced one 10-point ordinal scale of hostile political actions. To produce different numbers from the 10-point cooperation scale, I multiplied all of the newly derived hostility rank-ordered values by negative one. Thus, we can distinguish hostility from cooperation by the negative valence. I then ordered both the hostility and cooperation categories to produce two 10-point ordered scales. Second, I assigned the corresponding mean weighted value to each category transforming each of the 10-point ordinal scales into two interval-like scales. The hostility scale ranges from <0 to –100 and the cooperation scale ranges from >0 to +100. The newly derived interval-like scales are presented in Table 4.

These are the scores for the 17 expert judges who were willing to scale the events and assign weights. Interesting to note is that only one pair of categories switched ranks from the LDM ordinal scales. Note that event CC (allowing an opposition group to take power after an election) is ranked more cooperative than event CF (government ending nationwide state of emergency), contrary to the LDM ordinal cooperation scale.[21] As for hostility, the judges' ordinal scale is the same as the LDM conflict scale.

Table 4
Expert Judge Scale

Cooperation				Hostility			
Event	**(LDM Value)**	**Scale**	**Weights**	**Event**	**(LDMValue)**	**Scale**	**Weights**
CA	(100)	10	84.06	HD	(–10)	–1	–4.12
CD	(90)	9	77.17	HA	(–20)	–2	–7.88
CH	(80)	8	73.94	HG	(–30)	–3	–12.71
CC	(60)	7	69.47	HH	(–40)	–4	–28.82
CF	(70)	6	44.65	HC	(–50)	–5	–32.12
CI	(50)	5	41.23	HF	(–60)	–6	–60.18
CG	(40)	4	27.71	HJ	(–70)	–7	–63.65
CB	(30)	3	25.88	HI	(–80)	–8	–70.12
CJ	(20)	2	16.35	HB	(–90)	–9	–85.18
CE	(10)	1	9.94	HE	(–100)	–10	–90.71

To utilize the weights, match data corresponding to the value in parentheses with the new expert weight. For example, any event in the IPI data labeled 200 (Cooperation) is assigned a weighted value of 16.35. To calculate the weights, I took each weight assigned by each judge for each event and summed them. Then I divided by the number of judges (17) to produce a mean weight for each event. Then I ordered the events by mean weights from least to most. All of the hostile events were then assigned a negative sign.

These results add greater confidence in using the scales because both the IPI project investigators and the survey responses are similar. However, the new scales provide the researcher with more information about the magnitude of each event. For example, events CC and CA (the maximum average weighted value) are only separated by 14.59 points on a scale that ranges from +1 to +100. The judges tend to agree that events CC, CH, CD, and CA are all highly cooperative, yet do not differ much in intensity from one another. Specifically, the results tell us that in the expert judges' minds, "the resolution of an internal war" (CA) is only one and a fifth times more cooperative than "the allowance of an opposition group to take power after an election" (CC). We can also observe that "genocide" (event HE) is almost three times as hostile as a "government imposing a regional curfew" (event HC). Unlike ordinal scales, the newly derived interval-like scales allow us to differentiate the distance and intensity that one event is from another.

Next, I assess the reliability of the two newly derived scales. The results of the inter-judge correlations and composite reliability scores are presented in Tables 5 and 6. The most important measure for the scale's reliability, according to Zaninovich (1963, p. 64), is the measure of composite reliability, Cronbach's alpha in this case. The composite reliability for the hostility scale is .991 and the composite reliability for the cooperation scale is .976, showing that if we took another sample, the values would be almost the same. The composite reliability scores clear the +.800 benchmark measure set by Block (1961).

I want to point out that there is an influential judge in the sample. Note that judge number 17 correlates negatively with all the other judges in the sample for the cooperation events (Table 5).[22] I considered dropping the outlying judge from the sample,

Table 5
Correlation Matrix of Each of the Judge's Weighted Hostility Scale

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	.946	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	.911	.865	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	.955	.899	.927	1	-	-	-	-	-	-	-	-	-	-	-	-	-
5	.877	.806	.955	.896	1	-	-	-	-	-	-	-	-	-	-	-	-
6	.952	.881	.969	.966	.923	1	-	-	-	-	-	-	-	-	-	-	-
7	.821	.737	.886	.943	.841	.919	1	-	-	-	-	-	-	-	-	-	-
8	.866	.800	.947	.834	.963	.920	.757	1	-	-	-	-	-	-	-	-	-
9	.740	.665	.868	.693	.832	.777	.666	.837	1	-	-	-	-	-	-	-	-
10	.952	.914	.952	.981	.953	.962	.901	.897	.744	1	-	-	-	-	-	-	-
11	.716	.651	.678	.869	.701	.759	.899	.548	.379	.821	1	-	-	-	-	-	-
12	.876	.766	.949	.922	.933	.963	.938	.907	.818	.919	.751	1	-	-	-	-	-
13	.932	.941	.910	.954	.878	.950	.820	.866	.742	.951	.715	.876	1	-	-	-	-
14	.854	.728	.844	.912	.859	.851	.874	.740	.720	.902	.830	.847	.860	1	-	-	-
15	.895	.822	.926	.956	.961	.940	.922	.887	.737	.976	.861	.933	.896	.905	1	-	-
16	.865	.784	.845	.965	.842	.909	.969	.739	.602	.928	.954	.905	.864	.896	.947	1	-
17	.928	.890	.953	.882	.910	.927	.759	.938	.791	.917	.560	.856	.929	.775	.852	.744	1

Average inter-judge correlation = .865
Composite reliability score = .991

Table 6
Correlation Matrix of Each of the Judge's Weighted Cooperation Scales

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	.757	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	.942	.887	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	.747	.881	.823	1	-	-	-	-	-	-	-	-	-	-	-	-	-
5	.905	.818	.893	.842	1	-	-	-	-	-	-	-	-	-	-	-	-
6	.646	.655	.714	.763	.613	1	-	-	-	-	-	-	-	-	-	-	-
7	.745	.848	.783	.963	.769	.704	1	-	-	-	-	-	-	-	-	-	-
8	.821	.833	.862	.926	.788	.780	.917	1	-	-	-	-	-	-	-	-	-
9	.658	.892	.812	.784	.698	.458	.785	.787	1	-	-	-	-	-	-	-	-
10	.740	.922	.803	.922	.791	.683	.911	.892	.764	1	-	-	-	-	-	-	-
11	.957	.849	.974	.835	.904	.803	.793	.869	.702	.804	1	-	-	-	-	-	-
12	.855	.961	.950	.888	.905	.696	.822	.846	.818	.905	.927	1	-	-	-	-	-
13	.987	.783	.952	.708	.904	.622	.684	.776	.661	.730	.957	.880	1	-	-	-	-
14	.889	.738	.904	.704	.803	.476	.685	.742	.691	.717	.848	.855	.888	1	-	-	-
15	.816	.835	.864	.867	.746	.782	.890	.862	.703	.888	.866	.868	.785	.816	1	-	-
16	.751	.904	.898	.901	.788	.666	.865	.842	.902	.802	.842	.907	.741	.803	.838	1	-
17	-.082	-.053	-.141	-.385	-.092	-.347	-.422	-.331	-.257	-.109	-.121	-.059	-.040	-.123	-.333	-.384	1

Average inter-judge correlation = .702
Composite reliability score = .976

Overall Scores when hostility scale is added to cooperation scale:
Average inter-judge correlation: .784

but had no scientific justification for doing so.[23] I want an expert average score for each event type. Different scholars think about events differently and may categorize them in different ways. Therefore, I include all of the respondents in the sample.

DO THE NEW INTERVAL SCALES MATTER?

Ordinal scales impose a linear interval relationship between the events and their weights. The expert judges' scales demonstrate that neither the hostility nor the cooperation scale is linear. In particular, Figure 1 shows that the line depicting the new cooperation interval scale crosses above and then back below the straight line depicting the LDM (1998) ordinal scale. For every event cooperative event, the ordinal measure either over- or underrepresents the corresponding "true" measure. Similarly, Figure 2 shows that the LDM scale overrepresents the level of hostility for each of the events.

Given the reliability coefficients and the composition of the panel, I argue that the new scales are closer to their respective true scores than the ordinal ones. If we

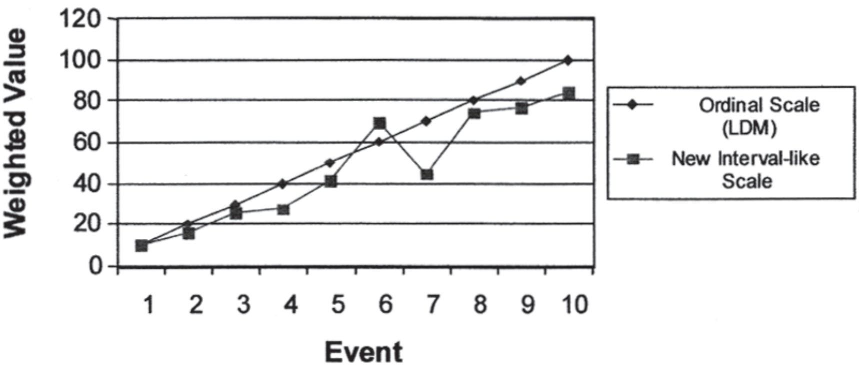


Figure 1. Ordinal Versus Interval Scale: Cooperation.

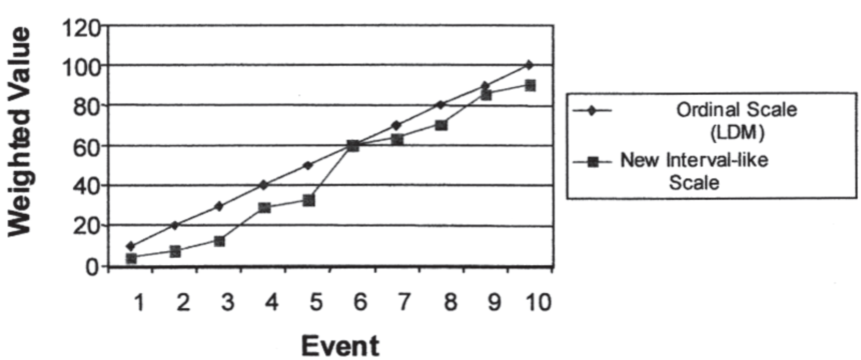


Figure 2. Ordinal Versus Interval Scale: Hostility.

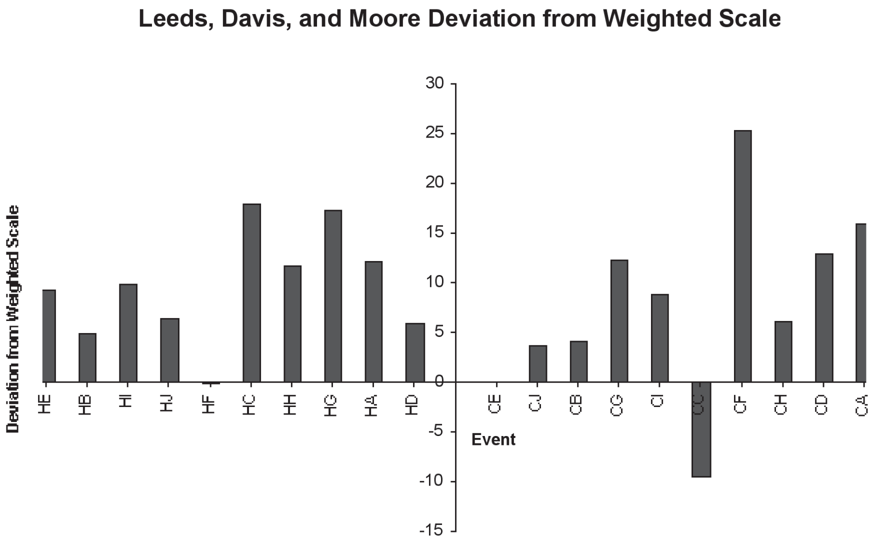


Figure 3. Leeds, Davis, and Moore Deviation from Weighted Scale.

assume for a moment that the weighted (interval-like) scale is the "true" score, we can illustrate that nonlinear, nonrandom measurement error contaminates the unweighted (ordinal) scale, which "*always* lead[s] to bias in OLS estimators" (Berry 1993, emphasis in the original).

Figure 3 depicts the events listed in Table 3 on the horizontal axis and the number of units, on the -100 to +100 H-C continuum, that the LDM scale deviates from the weighted scale. Overall, the unweighted data consistently overestimate some levels of cooperative behavior (event types CE, CJ, CB, CG, CI, CF, CH, CD, and CA) and violent behavior (event types HD, HA, HG, HH, HC, HJ, HI, HB, and HE) and consistently underestimate levels of cooperative (event CC) behavior. Though each event is consistently over- or underestimated by the same constant percentage, those constant percentages vary across event types. That is, the ordinal scale is nonlinearly related to the weighted scale (Figures 1, 2, and 3). Again, given the reliability of the collective scale and the composition of the panel that produced the interval-like scale, parameter estimates using the ordinal data are biased estimates.

However, because there are usually multiple variables in the model and event data must be aggregated by some unit of temporal aggregation, it is difficult to hypothesize the direction of the bias of the coefficient estimate. To begin, "if more than one variable is measured with error . . . coefficients from equations that ignore error in variance can be higher, lower, or even the same as the true coefficient" (Bollen 1989, 166–167). Second, aggregation provides another ingredient to the recipe for biased coefficient estimates. Freeman (1989, p. 92) argues that "variable aggregation has a tendency to simplify model forms and create spurious causation." Goldstein (1991, p. 207) adds that in studies of international conflict and cooperation, "over-aggregation in time series analysis" can "mask" the finding of reciprocating actions among actors.

While the effects of temporal aggregation on statistical estimates is still a bit fuzzy (Franzosi, 1995, Granger and Siklos, 1995, Marcellino, 1999, Hwang, 2000, Shellman, 2004a, Zellner and Montmarquette, 1971), most would argue that we must check the robustness of our findings over multiple units of temporal aggregation (Freeman, 1989).

Statistical Comparisons

To see how the new scale affects statistical results and the inferences researchers and policy-makers draw from analyses that use IPI event data, I estimate a series of models using weighted and unweighted data and compare and contrast the results.[24] Before discussing the results, I discuss issues concerning case selection, model specification, and temporal aggregation decisions. In addition, I briefly specify a set of hypotheses to test.

Hypotheses As an internal political conflict scholar, I am interested in explaining general patterns of state and dissident behavior. The IPI data are useful in analyzing such state-dissident behavioral interactions. In this particular analysis, I test competing hypotheses in the repression-dissent nexus. Specifically, I ask the two following questions: 1) is the effect of state repression on dissident behavior escalatory (positive) or deterrent (negative) and 2) is the effect of dissident hostility on state behavior escalatory or deterrent? *The backlash or escalatory hypothesis* states that increased hostility begets increased hostility, while the *deterrent hypothesis* states that increased hostility suppresses hostility. One should see Tilly (1978), Gurr (1970), Lee, Maline, and Moore (2000), Davenport (1995), Krain (2000), and Francisco (1995) for theoretical and empirical studies addressing these particular hypotheses, as well as additional hypotheses with regard to the repression and dissent literature. I choose to test the hypotheses in the contexts of Chile and Venezuela.

Case Selection I choose not to replicate published research for compound reasons. First, there is only one published study that uses the IPI data, and second, the researchers do not use the ordinal scales. Instead, Moore and Davis (1998) assign their own interval-like weights. In addition, Lee (2001) in his unpublished dissertation uses an entirely different set of subjective interval weights. Again, one purpose of constructing the new measures is to provide a *common*, valid, and reliable interval-like measuring stick for domestic political conflictual and cooperative behavior.

Instead of replicating a study, I empirically test the above hypotheses by analyzing two different IPI data sets. I do so for two reasons. First, I choose to examine both Chile (1983–1992) and Venezuela (1987–1992) because each country experienced civil conflict between domestic social actors and the ruling government. Thus, each case provides varying state and dissident actions and reactions to study. Second and most importantly, the comparative analysis rules out that the differing results across scales are artifacts of one particular case. That said, the purpose of the statistical analyses that follow is to examine whether or not parameter estimates are affected by different measures of the same concept. The purpose is *not* to show that Venezuelan state/dissident behavior is similar or dissimilar to Chilean state/dissident behavior.[25] Nonetheless, by performing the analyses in multiple contexts, I demonstrate that the sensitivity of results across scales is not specific to a single case.

Model Specification Similarly, I wish to show that the results that follow do not solely depend upon model selection. To do so, I choose to estimate parameters for general model specifications that are found frequently in the literature. I begin by estimating Richardson-type action- reaction systems of equations.[26] In short, I model state and dissident behavior as a function of one another's contemporaneous behavior and their own past behavior one period back. Next, following Goldstein (1992), I run a series of standard Vector Autoregression analyses (Goldstein and Freeman, 1990; Moore, 1995; McGinnis and Williams, 2001) and report the joint F-statistics. The models explain state behavior in terms of its own lagged values as well as the lagged values of dissident behavior and dissident behavior in terms of its own lagged values and the lagged values of state behavior.[27] We can test the escalatory and backlash hypotheses using both model specifications.

Aggregation Decisions Finally, I noted above that Freeman (1989, p. 92) and Goldstein (1991, p. 207) argue that the way in which we aggregate event data may affect the inferences we draw from regression analyses. To address this concern, I estimate both model types using daily, monthly, and "turns" time series data. A discussion of turn data is presented in the Appendix. I briefly discuss the affects of aggregation in these particular analyses in a section below.

Estimation

To avoid spurious results associated with regressing one nonstationary time series on another nonstationary time series, each time series in the model must be stationary. Using augmented Dickey-Fuller tests, I determined that each of the series (S_t and D_t) for each scale (ordinal and interval) for each temporal unit of aggregation (days, months, and turns) is stationary or I(0).[28] The simultaneous structural equation models used to analyze the daily and monthly series are estimated using two-stage least squares (2SLS), whereas each equation specified to analyze the turn data is estimated by OLS on its own.[29] Since an unrestricted vector autoregressive model is a system of unrestricted reduced form equations, OLS estimates will be consistent.

Interval-like (Weighted) and Ordinal (Unweighted) Parameter Comparisons

The results are reported in Table 7 and Table 8. In general, one should compare the weighted results to the unweighted results *within* each case and temporal unit of aggregation. The bold face entries represent changes across scales.[30] I begin by highlighting the differences within each model specification and then discuss the general findings and implications. Table 7 (A and B) reports the results for the structural equation models. To answer the first question regarding the effect of repression on dissident behavior, we observe the sign and statistical significance of the coefficient for S_t (days and months) or S_{t-1} (turns), where D_t is the dependent variable. To answer the second question regarding the effect of dissident behavior on state behavior, we observe the coefficient on D_t, where S_t is the dependent variable. Positive statistically significant coefficients provide support for the backlash hypothesis, whereas negative statistically significant coefficients provide support for the deterrent hypothesis. We conclude that there is no relationship between the variables if the coefficient is not statistically significant. In the discussion of results, I will concentrate on comparing the statistical significance levels of the coefficients across the scales. Then, I will substantively interpret some of the parameters.

Concentrating on the upper left quadrant of Table 7A, as we move from the unweighted to the weighted results, we see for the *daily* Chilean analyses the significance levels increase so that coefficients become insignificant. For example, the coefficient for S_t is significant at the .10 level in the unweighted analysis, but loses its statistical significance in the weighted analysis. The same general pattern is true for S_{t-1} and D_{t-1}. In particular, the unweighted coefficient for D_{t-1} is much larger than the weighted coefficient and is statistically significant at the .001 level. The smaller weighted coefficient is not statistically significant. Overall, using the ordinal scale we conclude that both past state behavior and past dissident behavior drive future dissident behavior, while we would conclude that dissident behavior is not driven by state or dissident behavior using the weighted analyses.

As for explanations of state behavior, we would conclude using the unweighted scale that both dissident and state behavior drive state behavior but when examining the weighted scale, we conclude that only dissident behavior drives state behavior. When we rely on the interval-like data, we conclude based on the positive coefficient on Dt, Chilean dissent increases Chilean repression.[31] We can make no additional inferences based on the analyses.

Examining the upper middle quadrant of Table 7A, we see that the weighted *monthly* Chilean results do not differ that much from the unweighted results. Only the statistical significance for the coefficient S_{t-1} changes. Moving from the unweighted results to the weighted results, the coefficient S_{t-1} achieves statistical significance. By analyzing the ordinal data only, we miss the relationship between the state's past behavior and its current behavior. As for the upper right quadrant, the *turns* analyses do not differ across scales.

Examining the Venezuelan results (Table 7B), the only significant change across scales occurs in the *daily* aggregated analysis. The coefficient for D_t achieves statistical significance above the .001 level in the weighted analysis. Sticking to the standard level of significance (.05) and using the ordinal data, researchers may neglect that previous state behavior drives future state behavior. Moving to the turns data someone adhering to a .10 significance level, using the weighted scale, aggregated by turns, may report a significant relationship between previous state behavior and future dissident behavior. However, no such relationship holds using the ordinal turn data. Drawing inferences from the ordinal data, we miss alarms. That is, we reject the possibility of particular relationships between states and dissidents using the unweighted data.

Substantively, interpreting the weighted daily D_t coefficient (.169) using the standard .05 level of significance in Venezuela, we find that on any given day, holding previous state behavior constant, if dissidents begin to riot perhaps causing more than 10 deaths (-60.18 on the weighted scale), that the Venezuelan state, in response, increases its hostility level, on average, by almost 11 points (10.76 on the weighted scale). If we suppose for a moment that the state previously acted at -60.18 (or equal to the dissidents), we would predict that the state responds to the dissident riots with increased government violence killing hundreds (-70.94 or -60.18 + 10.76).[32] We

Table 7
Structural Equation Models

A: Dissident-State Political Interactions in **Chile, 1983–1992**

	Daily Aggregation (2SLS)				**Monthly Aggregation (2SLS)**				**Turns Aggregation (Sequential Models - OLS)**			
	Unweighted		*Weighted*		*Unweighted*		*Weighted*		*Unweighted*		*Weighted*	
Variable	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt
S_t	–	**1.14†** **(.700)**	–	**1.20** **(.795)**	–	.274 (.554)	–	.051 (.452)	–	–	–	–
D_t	.178** (.063)	–	.415** (.172)	–	1.16 (.968)	–	.936 (.850)	–	.482*** (.075)	–	.617*** (.076)	–
S_{t-1}	**.028†** **(.017)**	–	**.015** **(.019)**	–	**.144** **.173**	–	**.226†** **(.129)**	–	.294*** (.044)	.056* (.029)	.220*** (.045)	.151*** (.029)
D_{t-1}	–	**.193***** **(.040)**	–	**.052** **(.045)**	–	.090 (.131)	–	.123 (.115)	–	.225*** (.049)	–	.112* (.050)
N	3556		3556		116		116		436		436	

Table 7
Structural Equation Models (continued)

B: Dissident-State Political Interactions in **Venezuela, 1987–1992**

	Daily Aggregation (2SLS)				**Monthly Aggregation (2SLS)**				**Turns Aggregation (Sequential Models - OLS)**			
	Unweighted		*Weighted*		*Unweighted*		*Weighted*		*Unweighted*		*Weighted*	
Variable	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt
S_t	–	-.659 (.459)	–	-.262 (.605)	–	-.241 (1.60)	– –	-1.35 (3.15)	–	–	–	–
D_t	**.037†** **(.021)**	–	**.169**** **(.037)**	–	.644 (1.77)	–	.737 (.998)	–	.364*** (.097)	–	.433*** (.086)	–
S_{t-1}	.080*** (.022)	–	.055** (.022)	–	.123 (.167)	–	.178 (.198)	–	.104 (.089)	**.141** **(.090)**	.131 (.084)	.168† (.100)
D_{t-1}	–	.622*** (.025)	–	.421*** (.052)	–	.077 (.144)	–	.218 (.329)	–	.096 (.101)	–	.104 (.104)
N	2115		2115		70		70		118		118	

S and D are state behavior and dissident behavior respectively. The subscript indicates the time of action. The dependent variables are across the top and the independent variables appear in column one. All models are run with constants but those coefficients do not appear in the table. The differences across scales are highlighted. I ran the models using both two- and three-stage least squares. I report the two-stage results. The three-stage results are almost identical to the two-stage results with no threats to inferences drawn from the analyses.

*** Significant at the .001 level   * Significant at the .05 level

** Significant at the .01 level † Significant at the .10 level

Table 8
Vector Autoregression (VAR) Analyses
Significance of F-Statistics (Granger Causality) Results

A: Dissident-State Political Interactions in **Chile, 1983–1992**

	Daily Aggregation				**Monthly Aggregation**				**Turns Aggregation**			
	Unweighted		*Weighted*		*Unweighted*		*Weighted*		*Unweighted*		*Weighted*	
Variable	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt
Dt	–	–	–	–	–	–	–	–	32.82***	–	20.81***	–
St_1	**2.87****	**2.23****	**1.64†**	**1.34**	**1.36**	.36	**2.65***	.22	6.37***	**.87**	8.04***	**8.04*****
Dt_1	**1.82†**	55.73***	**2.12***	46.7***	**2.35†**	1.97†	**2.56***	2.04†	–	**7.94*****	–	**2.21†**
# of lags	10	10	10	10	4	4	4	4	3	3	3	3
N	3547		3547		113		113		433		433	

Table 8
Vector Autoregression (VAR) Analyses
Significance of F-Statistics (Granger Causality) Results (continued)

B: Dissident-State Political Interactions in **Venezuela, 1987–1992**

	Daily Aggregation				**Monthly Aggregation**				**Turns Aggregation**			
	Unweighted		*Weighted*		*Unweighted*		*Weighted*		*Unweighted*		*Weighted*	
Variable	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt	St	Dt
Dt	–	–	–	–	–	–	–	–	2.87	–	4.15***	–
St_1	8.76***	**1.37**	6.15***	**2.56****	.73	.52	.80	.35	1.49	**2.02**†	1.36	**2.26***
Dt_1	4.03***	221.74***	7.50***	78.50***	.55	1.07	1.24	.67	–	1.27	–	1.17
# of lags	8	8	8	8	8	8	8	8	6	6	6	6
N	2107		2107		63		63		113		113	

S and D are state behavior and dissident behavior respectively. The subscript indicates the time of action. The dependent variables are across the top and the independent variables appear in column one. All models are run with constants but those coefficients do not appear in the table. The differences across scales are highlighted. Each cell contains the joint-F statistic for each set of lagged variables. The lag lengths were specified using the Akaike and Schwartz criterion. It just so happens that the weighted and unweighted lag specifications are the same for each temporal aggregation. The way the data are stacked in turns aggregations, the first lag specified in an equation where the dependent variable is S_t is D_t (see appendix).

*** Significant at the .001 level * Significant at the .05 level

** Significant at the .01 level † Significant at the .10 level can make no such inference from the ordinal data if we are adhering to the standard level of statistical significance.

Alternatively, if we choose a significance level of .10 using the ordinal data, we can substantively interpret the coefficient estimate, but the estimate is biased. To demonstrate, let's assume the same scenario as presented above but let's measure the events using the ordinal scale. Substantively, the same scenario suggests that the dissident riots only produce a 2.2 point behavior increase on the -100 to +100 ordered scale. The state would behave at or about -62.2. This value is not enough to pass the constant ten-point interval distance between two categories. Thus, the unweighted model would suggest that state behavior does not increase to the next scaled level. In contrast, the weighted analyses showed a jump of two categories (-63.65 and -70.12) in one particular instance.[33] In short, different inferences are drawn using the ordinal scale as opposed to using the interval-like scale.

Moving on to the VAR analyses, Table 8 presents the statistical significance levels of the F-tests for each set of lagged variables (indicating a lagged correlation with the dependent variable in the equation). In the VAR analyses, we cannot directly assess the statistical significance of individual lagged coefficients (Gujarati, 1995, pp. 749–750).[34] Instead, we must rely on impulse response functions and joint F-statistics to draw substantive and statistically significant inferences. In this brief piece, I present only the joint-F statistics.[35] A statistically significant F allows us to conclude that a particular set of lagged variables (e.g., D_{t-1}, $D_{t-1, \ldots}$ D_{t-n}) explain the variance in the dependent variable (e.g., S_t). For example, does past dissident behavior drive current state behavior and does past state behavior drive current dissident behavior?

A quick glance at the boldface entries of Table 8A and Table 8B shows that the results across scales change quite a bit for the Chilean analyses, while only altering two F-statistics in the Venezuelan analyses. For the *daily* temporal unit of aggregation, the lagged Chilean state F (F = 1.34) for the weighted analyses is not statistically significant, meaning that previous days of state activities do not drive today's dissident activities. A positive relationship is supported using the ordinal data. In contrast, using a standard level of significance (.05) we observe that previous days of dissident activities (F = 2.12) impact today's dissident behavior in the interval-like analyses as opposed to the ordinal analyses. Thus, if we were to use the ordinal data, we would draw a false conclusion about a positive relationship between state and dissident behavior in the former instance and a miss a relationship between past dissident behavior and current dissident behavior in the latter instance. We find similar patterns across the scales in the *monthly* and *turn* cases. Both previous state and dissident behavior is only found to predict current state behavior in the Chilean interval-like *monthly* data, as opposed to the ordered *monthly* data. Finally, in the *turn* analyses, we see that Chilean state behavior only effects dissident behavior when the interval-like data are used.

In Venezuela, the interval-like data produce two additional statistically significant parameters (at the .05 level) than their ordinal counterparts. In the *daily* analyses using the interval-like scale, lagged state behavior effects current dissident behavior. In contrast, the ordinal data produce no relationship at the daily level. We find the same pattern at the turn level of aggregation. With respect to the VAR analyses, inferences drawn from the interval data vary substantially from the inferences drawn from the ordinal data.

Overall, we see that the interval-like scale produces different results than those produced using the LDM scale in both cases, in both models, and in each temporally aggregated unit. Many significance levels increase and decrease across scales and point estimates also fluctuate across the scales. Why do the estimates differ in magnitude and in statistical significance levels? The LDM scale over and under estimates domestic political actors' behavior. When the events are aggregated, the over- and underrepresented events are added together. The coefficient estimate will sometimes be smaller and often not statistically significant from zero or larger and statistically significant depending on the amount of conflict and cooperation events and the variety of event types aggregated together. Therefore, the OLS estimates using the LDM scale are biased and inefficient.

In addition, there is a difference with respect to how much measurement matters across cases. While there is evidence of parameter differences across the scales in Venezuela, there are less instances than in Chile. Particularly, there are no differences in Chile at the monthly level of aggregation. Speculation draws attention to the difference in sample size in the monthly data. Venezuela only has 70 months of data to analyze in the structural equation models and only 63 in the VAR analyses. As the sample size increases, a given coefficient is more likely to be significant. Perhaps the small sample size is effecting the monthly Venezuelan parameter estimates.[36] None of these estimates, weighted and unweighted, are statistically significant. An additional hypothesis is that there are more events like HF, CE, and CJ in Venezuela than Chile. The ordinal measures on these events are close to the interval measures, which would introduce less bias. However, a look at the raw data indicates the percentage of these event types is similar across cases. Unfortunately, I have no definitive answer on why the scales seem to matter more in Chile than Venezuela. Nonetheless, we do find significant parameter changes across the scales in both cases. Thus, the results are not just an artifact of a single case. Moreover, changes across temporal units of aggregation and model specifications imply that the interval-like scales affect statistical inference. Before concluding the paper, I briefly address the issue of temporal aggregation.

Temporal Aggregation and Threats to Inference

Do aggregation decisions affect inferences? While I cannot completely address the effects of temporal aggregation on parameter estimation in this article (See Shellman 2004a for additional work on temporal aggregation and inference), I briefly make some remarks about how these particular coefficient estimates are affected by aggregation decisions. For this purpose, I will concentrate on the *weighted* structural equation models. In most political science research, scholars are interested in testing the direction of a particular relationship. For example, we tested competing hypotheses in the repression and dissent literature as to the effects of state behavior on dissident behavior and dissident behavior on state behavior. Though none of the coefficients switch signs across units of aggregation, there are instances of differing levels of statistical significance. Examining the Chilean analyses, we see that the dissidents (.151***) only respond to the state when the data are aggregated by turns.[37] Using the turns data, we conclude that dissident behavior is positively associated with state behavior. In other words, as the state becomes more cooperative, the dissidents become more cooperative, and as the state becomes more hostile, the dissidents become more hostile. Therefore, we conclude that state repression escalates dissent. However, if we were to study this question using a daily (1.21) or monthly (-.017) temporal unit of aggregation, we would conclude that there is no relationship between state behavior and dissident behavior. Using the structural equation analyses for the Venezuelan case, we conclude that state repression escalates dissent only at the turn level of aggregation.

Examining the Chilean and Venezuelan results in reference to the second question, we conclude at both the daily and turn levels that a positive relationship exists between dissident behavior and state behavior, but at the monthly level there is no statistically significant relationship. The message is that parameter estimates are affected by aggregation decisions. Therefore, one should examine the sensitivity of results across various units of aggregation when assessing the tenability of knowledge claims founded in time series econometric studies. More weight should be given to results that hold across multiple levels of aggregation. Regardless, first and foremost, I would encourage scholars to consult theory when choosing a temporal unit of aggregation.[38]

CONCLUSION

This paper sketches the logic and design behind measuring the IPI events data on an interval scale. I follow Azar and Sloan (1975), Goldstein (1992), and Moore and Lindstom (1996) in conducting a survey of experts to collectively scale and weight events data on separate cooperation and hostility dimensions. The survey produces two valid and reliable interval-like scales for measuring cooperative and hostile actions. The results of this project are significant in at least three important regards. First, there now exist two valid reliable interval-level scales for measuring cooperative and hostile IPI events. Carmines and Zeller (1979, p. 16) state "for an indicator to be useful in social science research, it must lead to quite consistent results on repeated measurements [reliability] and reflect its intended theoretical concept [validity]." By collectively scaling the IPI events, we are more confident that the scales are valid. Moreover, the reliability coefficients illustrate that the measures are consistent. Based on the reliability coefficients and the panel composition, the new scales represent measures closer to the true scores.

Second, utilizing the weights allows for more comparable results across countries and methods. In the past, Lee (2001) and Moore and Davis (1998) each created their own interval scales. The new interval-like scale provides researchers with a set of standard weights derived by the group of experts. By using the standardized measures, we eliminate inconsistencies that arise from the use of different measures. The weighted-scales provide common bases of measurement for political interactions. I encourage scholars who use the IPI data in statistical models to use the weights in order to eliminate the inconsistencies associated with random assignments of weighted values to the ordered categories.

Third, the interval-like scales are more appropriate than the previous ordinal scales for OLS regression analysis. Lodge (1982, p. 71) argues that "Regression coefficients computed from categorical [ordinal] judgments are indeterminate." The interval-like scales convey greater distinctions among the conflict–cooperation events and allow for meaningful interpretation of regression coefficients.

The project derives new valid and reliable cooperation-hostility weights that effectively and efficiently quantify subjective measurement on an interval-level scale. In the words of King, Keohane, and Verba (1994, pp. 27–28), this project maximizes the validity and reliability of the IPI data and enables research to be replicable and comparable across cases. [39]

NOTES

1. See Leeds, Davis, and Moore (1996) for a full description of the project and/or www.garnet.acns.fsu.edu/whmoore/ipi/ipi.html.
2. For example, events, A, B, C, and D are scaled such that A = 10, B= 20, C=30, and D=40.
3. There exist two different non-standard IPI interval scales in the literature. I discuss each below.
4. Of course, there are others such as PANDA—see www.wcfia.harvard.edu/ponsacs/panda.htm
5. See Leeds, Davis, and Moore (1996) or Davis, Leeds, and Moore (1998). The period of time that the data covers varies from case to case. Some cases begin in 1979 or 1980. However, many of the cases cover the period 1983–1992. Countries included in the project are Argentina, Belgium, Brazil, Chile, Colombia, Hungary, Indonesia, Mexico, Nigeria, Pakistan, South Korea, Venezuela, Zaire, and Zimbabwe.
6. Note that independent variables may be dichotomous. Some assume that noninterval data are actually interval when including them in regression equations. Others often assume that ordinal data are quantitative qualifying them for regression analyses. Berry (1993, p. 47) argues that to be appropriate as explanatory variables in a regression model, "ordered discrete variables must be quantitative," and that not all ordinal data are necessarily quantitative. Finally, to appear as dependent variables, they must be "continuous" (Berry, 1993, p. 12). I take steps to increase our confidence that the quantitative, continuous, and interval assumptions are met.
7. Of course there are alternative views. See Bohrnstedt and Carter (1971) and Labovitz, (1970). A middleground perspective is offered by Asher (1976).
8. In this exercise, Moore and Davis convert 20 ordered IPI categories into 15 interval-like VICDP categories.
9. See Berry (1993, p. 49–57). The true score may never be known, however we increase our confidence that our indicators are equivalent to their true scores by increasing reliability and validity of our measures.
10. The research question changes from one about levels, intensity and/or change of behavior to one about frequency.
11. I do not sample just any political scientist or any international relations faculty or any conflict scholar for that matter. The panel comes from those who specifically study *internal* conflict.
12. This strategy is known as the Multitrait-Multimethod (MTMM) strategy.
13. Note that the sample selected for this endeavor does not include Leeds, Davis, or Moore, the creators of the original IPI scales nor does it include myself. The doctoral student is not from my university. The respondents shall remain anonymous.
14. With the exception of one scholar, the experts had multiple publications in these areas.
15. "Eighteen scholars and practitioners of international relations" scaled the events for the COPDAB dataset (Azar and Sloan, 1975), while eight faculty of the School of International Relations at the University of Southern California weighted the WEIS events data (Goldstein, 1992).
16. That is, roughly the same percentage of political scientists and sociologists, quantitative and qualitative scholars, and assistant, associate, and full professors make up the respondents as those sampled.
17. I pretested the directions and procedure by administering the survey to faculty and graduate students

at my university. Upon completion of the procedure, I received comments and suggestions on how the directions and forms could be presented more clearly. Before sending out the "judging-packet," I made the necessary changes and improved the original packet.

18. Note that some events exemplify actions taken by the state towards a dissident group, while others exemplify actions taken by a dissident group towards the state. I attempted to split the 20 events up between these two types of scenarios where possible.
19. Note that later in the process when judges assign weights to the events they can assign the same value to multiple categories if they desire.
20. Robinson, et al. (1974, p. 251) state that "of all the indices that have been proposed [to assess reliability], however, probably none combines the simplicity with amount of information contained as well as the inter-item correlation matrix." However, I do not wish to assess how well one item correlates with another, instead I wish to determine the strength of the relationship between each pair of judges' scales. Therefore, I construct an inter-judge correlation matrix.
21. Note that many judges weighted event CC near the upper limit of the cooperation scale (100), while others weighted it at or well below the mean.
22. The judge only negatively correlates with the other judges on the cooperation scale. He/she correlates positively and highly with the other judges on the hostility scale.
23. I did in fact compute the results excluding the judge to see exactly how much influence he/she had on the results. The categories remained in their ordered position and some weights only changed up or down a few points. One judge in seventeen did not have a drastic effect on the newly derived scale. However, the average inter-judge correlation increased to .823. That said, the composite reliability remains high, the average inter-judge correlation is above the inferential rule of .70, and the weighted scale remains relatively unchanged. Based on the facts that I have no scientific grounds for excluding the judge, the results remain relatively unchanged, and validity is increased, I chose to include the outlier.
24. In this particular set of analyses, the ordinal measures, -100, -90, . . . , 90, 100 were used to increase the comparability of estimates. This favors not finding a difference between ordinal and interval scales. In a separate set of analyses, I examine the raw measures from the IPI data set whose categories range from -1000 to +1000—see Leeds, Davis, and Moore (1996). The results obtained from the raw measures are consistent with the ones reported in the paper. Slight differences exist, with respect to magnitude of the coefficients, but significance levels remain the same. The raw measures do not change our inferences.
25. For a specific theory, test, and comparison of the Venezuelan and Chilean cases, see Shellman (2004b).
26. See Richardson (1960), Intriligator and Brito (1989), and Rajmaira and Ward (1990) for more on theoretical and empirical arms race models that describe the Richardson equations in more detail.
27. This specification is the reduced form of a system specified as a function of both contemporaneous and lagged terms (see Goldstein and Freeman, 1990, 67–70).
28. Since the levels are stationary, the lags are stationary. However, I dismiss the possibility of fractionally integrated series (see Box-Steffensmeier and Smith, 1998).
29. The daily and monthly simultaneous equation models are exactly identified nonrecursive models and 2SLS is an appropriate estimation strategy (Gujarati, 1995, p. 700). I also run three-stage least squares. The results (not reported here) are consistent with the two-stage results. The two-equation turn models do not share contemporaneous endogenous terms. If we assume that the errors are uncorrelated, the model is recursive and each equation can be estimated on its own (Bollen, 1989, pp. 95–98,104). If we believe the errors are correlated, we can use a three-stage estimator or a FIML technique. When checking the correlation between the residuals across equations, I found low values. Thus, I use OLS to estimate each equation on its own.
30. The criteria for bold face type are (1) a change from statistical significance (<.10) to non-statistical significance (>.10) or vice versa, (2) a change from .10 level of statistical significance to a level < .05 or vice versa, and/or (3) a sign change that is statistically significant.
31. A positive coefficient indicates a positive linear relationship such that as dissident cooperation increases, state cooperation increases and as dissident cooperation decreases (i.e., hostility increases), state cooperation decreases (i.e., hostility increases).
32. See Table 1 and corresponding weights. I do not consider the coefficient on S_{t-1} in this substantive explanation.

33. Though 11 points is not always two categories on the interval-like scale, the interval-like coefficient is greater than its ordinal counterpart, which at the .05 level is not statistically significant from zero.
34. High multicollinearity among the independent variables inflates the standard errors biasing t-statistics.
35. The impulse response functions communicate the direction of the relationship. However, due to space, I present only the F-statistics. The F tests sufficiently convey the point made in the paper that the results differ across the scales.
36. In all cases, the N is smaller in Venezuela than in Chile.
37. However, the VAR analyses report statistically significant relationships at the daily level.
38. For example, it makes sense to analyze budgets made every year in annual units. Alternatively, it is a bit fuzzier to construct appropriate temporal units of aggregation for state and dissident behavior. However, I argue that if we want to study action-reaction behavior (as I wish to do) we should study behavior in "turn-taking" sequences. Alternatively, if we want to study the timing of one's actions (or time between actions), turns is not the best temporal unit of aggregation.
39. Lastly, Doug Bond and I have discussed the possibility of mapping the IDEA codes to the IPI scale. The IDEA project has sought to produce data that can be converted to myriad scales including IPI (see http://www.vranet.com/IDEA/). Therefore, the resulting interval-like scale produced in this study may also be relevant to the IDEA data project.
40. Note that events data essentially codes each "move" made by the dissidents toward the state and every move made by the state towards the dissidents. One may choose to use sums in contrast to means.

REFERENCES

Asher, H. (1976). *Causal Modeling*. Sage University Papers Series on Quantitative Applications. Beverly Hills, CA: Sage.

Azar, Edward E. (1982). *The Codebook of the Conflict and Peace Data Bank* (COPDAB). Typescript, University of North Carolina at Chapel Hill.

Azar, Edward E. and Thomas Sloan (1975). *Dimensions of Interactions*, Pittsburgh: University Center for International Studies, University of Pittsburgh

Bakeman, Roger and John M. Gottman (1997). *Observing Interaction: An Introduction to Sequential Analysis*, 2nd ed., Cambridge: Cambridge University Press.

Berry, William D. (1993). *Understanding Regression Assumptions*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-092. Newbury Park, CA: Sage.

Berry, William D. and Stanley Feldman (1985). *Multiple Regression in Practice*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-050. Newbury Park, CA: Sage.

Block, Jack (1961). *The Q-sort Method in Personality Assessment and Psychiatric Research*, Springfield: Charles C. Thomas Publisher.

Bohrnstedt, G. W. and T. M. Carter (1971). "Robustness in Regression Analyses," in H. L. Costner, ed., *Sociological Methodology*, San Francisco: Jossey-Bass.

Bollen, Kenneth A. (1989). *Structural Equations with Latent Variables*, New York: John Wiley and Sons.

Box-Steffensmeier, Janet M., and Renée M. Smith (1998). "Investigating Political Dynamics Using Fractional Integration Methods." *American Journal of Political Science*, Vol. 42, No. 2 (April), pp. 661–689.

Campbell, D. T. and Fiske, D. W. (1959). "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin,* Vol. 56, pp. 81–105.

Carmines, Edward D. and Richard A. Zeller (1979). *Reliability and Validity Assessment*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-017, Newbury Park, CA: Sage.

Davis, David R., Brett Ashley Leeds, and Will H. Moore (1998). "Measuring Dissident and

State Behavior: The Intranational Political Interactions (IPI) Project." Presented at the Workshop on Cross-National Data Collection. Texas A&M University, November 21.

Davenport, Christian (1995). "Multi-Dimensional Threat Perception and State Repression: An Inquiry into Why States Apply Negative Sanctions." *American Journal of Political Science* Vol. 38, No. 3, 683–713.

Dixon, William J. (1988). "The Discrete Sequential Analysis of Dynamic International Behavior," *Quality and Quantity*, Vol. 22, pp. 239–254.

Francisco, Ronald A. (1995). "The Relationship between Coercion and Protest: An Empirical Evaluation in Three Coercive States." *Journal of Conflict Resolution*, Vol. 39, pp. 263–382.

Franzosi, Roberto (1995). *The Puzzle of Strikes: Class and State Strategies in Postwar Italy*, New York: Cambridge University Press.

Freeman, John R. (1989). "Systematic Sampling, Temporal Aggregation, and the Study of Political Relationships." *Political Analysis*, Vol. 1, pp. 61–88.

Goldstein, Joshua S. (1991). "Reciprocity in Superpower Relations: An Empirical Analysis." *International Studies Quarterly*, Vol. 35, pp. 195–209.

Goldstein, Joshua S. (1992). "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution*, Vol. 36, pp. 369–385.

Goldstein, Joshua S. and John R. Freeman (1990). *Three-Way Street: Strategic Reciprocity in World Politics*, Chicago: University of Chicago Press.

Goldstein, Joshua S. and John R. Freeman (1991). "U.S.–Soviet–Chinese Relations: Routine Reciprocity, or Rational Expectations?" *American Political Science Review*, Vol. 85, No. 1, pp. 17–36.

Granger, C. W. J. and Pierre L. Siklos (1995). "Systematic Sampling, Temporal Aggregation, Seasonal Adjustment, and Cointegration: Theory and Evidence," *Journal of Econometrics*, Vol. 66, pp. 357–369.

Gujarati, Damodar N. (1995). *Basic Econometrics*, New York: McGraw-Hill, Inc.

Gurr, Ted Robert (1970). *Why Men Rebel*, Princeton: Princeton University Press.

Hwang, Soosung (2000). "The Effects of Systematic Sampling and Temporal Aggregation on Discrete Time Long Memory Processes and their Finite Sample Properties," *Econometric Theory*, Vol. 16, pp. 347–372.

Intriligator, Michael D. and Dagobert L. Brito (1989). "Richardsonian Arms Race Models," in M.I. Midlarsky, ed., *Handbook of War Studies*, Ann Arbor: Michigan University Press, pp. 219–236.

King, Gary, Robert O. Keohane, and Sidney Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, NJ: Princeton University Press.

Krain, Matthew (2000). *Repression and Accommodation in Post-Revolutionary States*, New York: St. Martin's.

Labovitz, S. (1970). "The Assignment of Numbers to Rank Order Categories." *American Sociological Review*, Vol. 35, pp. 515–524.

Lee, Chris (2001). *Protest and Repression in Latin America: A Synthetic Model*, unpublished dissertation, University of California, Riverside.

Lee, Chris, Sandra Maline, and Will H. Moore (2000). "Coercion and Protest: An Empirical Inquiry Revisited." In C. Davenport, ed., *Paths to State Repression: Human Rights and Contentious Politics in Comparative Perspective*, Boulder: Rowman and Littlefield, pp. 127–147.

Leeds, Brett Ashley, David R. Davis, and Will H. Moore with Christopher McHorny (1996). Intranational Political Interactions (IPI) project *Codebook*. Unpublished manuscript. Florida State University: Tallahassee, FL.

Lodge, Milton (1982). *Magnitude Scaling of Social Psychological Judgments*, Beverly Hills,

CA: Sage Publications, Series on Quantitative Applications in the Social Sciences.
Manheim, Jarol B. and Richard C. Rich (1995). *Empirical Political Analysis: Research Methods in Political Science*, White Plains, NY: Longman.
Marcellino, Massimiliano (1999). "Some Consequences of Temporal Aggregation in Empirical Analysis," *Journal of Business and Economic Statistics*, Vol. 17, No. 1, pp. 129–136.
Marks, L. E. (1974). *Sensory Processes: The New Psychophysics*, New York: Academic Press.
Marlin-Bennett, Renee, Alan Rosenblatt, and Jianxin Wang (1991). "The Visible Hand: The United States, Japan, and the Management of Trade Disputes." *International Interactions*, 17, pp. 191–213.
McGinnis, Michael D. and John T. Williams (2001). *Compound Dilemmas: Democracy, Collective Action, and the Superpower Rivalry*, Ann Arbor: University of Michigan Press.
Moore, Will H. (1995). "Action-Reaction or Rational Expectations: Reciprocity and the Domestic International Conflict Nexus During the 'Rhodesia Problem.'" *Journal of Conflict Resolution*, Vol. 39, No. 1, pp. 129–167.
Moore, Will H. (1998). "Repression and Dissent: Substitution, Context, and Timing." *American Journal of Political Science*, Vol. 42, No. 3, pp. 851–873.
Moore, Will H. (2000). "The Repression of Dissent: A Substitution Model of Government Coercion." *Journal of Conflict Resolution*, Vol. 44, pp. 107–127.
Moore, Will H. and David R. Davis (1998). "Ties that Bind? Domestic and International Conflict Behavior in Zaire," *Comparative Political Studies*, Vol. 31, pp. 45–71.
Moore, Will H. and Ronny Lindstrom (1996). *The Violent Intranational Conflict Data Project (VICDP) Codebook*, unpublished manuscript, Riverside: University of California.
Rajmaira, Sheen and Michael D. Ward. (1990). "Evolving Foreign Policy Norms: Reciprocity in the Superpower Triad." *International Studies Quarterly*, Vol. 34, No. 4, pp. 457–475.
Richardson, Lewis F. (1960). *Arms and Insecurity*, Pittsburgh: The Boxwood Press and Quadrangle Books.
Robinson, John P., Jerrold G. Rush and Kendra B. Head (1974). "Criteria for an Attitude Scale." In Maranell, Gary M., (ed.), (1974). *Scaling: A Sourcebook for Behavioral Scientists*, Chicago: Aldine Publishing Company.
Schumacker , Randall E., Michael P. Monahan, and Robert E. Mount (2002). "A Comparison of OLS to LTS and MM Robust Regression in S-PLUS. " Paper presented at the Southwest Educational Research Association 25th annual meeting. February 14–16, 2002, Austin, Texas.
Shellman, Stephen M. (2004a). "Time Series Intervals and Statistical Inference: The Effects of Temporal Aggregation on Event Data Analysis." *Political Analysis*, Vol. 12, No. 3, pp. 97–104.
Shellman, Stephen M. (2004b). "Taking Turns: A Theory and a Model of Government-Dissident Interactions." Unpublished dissertation. Florida State University: Tallahassee, FL.
Shinn, A., Jr. (1974). "Relations between Scales," in H. M. Blalock, Jr. (ed.), *Measurement in the Social Sciences: Theories and Strategies*, Chicago: Aldine.
Sloan, T. (1975). "The Development of Cooperation and Conflict interaction Scales: Advance in the Measurement and Analysis of Events Data." In E. E. Azar and J. D. Ben-Dak (eds.), *Theory and Practice of Events Research: Studies in Inter-Nation Actions and Interactions*, pp. 29–39, New York: Gordon and Breach.
Tilly, Charles (1978). *From Mobilization to Revolution*, New York: Random House.
Vincent, J. E. (1979). *Project Theory*, Lanham: University Press of America.
White, Benjamin W. and Eli Saltz (1974). "Measurement of Reproducibility." In Gary M. Maranell, (ed.), 1974. *Scaling: A Sourcebook for Behavioral Scientists*, Chicago: Aldine Publishing Company.
Wilson, Thomas (1971). Critique of Ordinal Variables. Chap. 24 in H. M. Blalock, ed., *Causal*

models in the social sciences, Chicago: Aldine.
Zaninovich, George M. (1963). "Measuring the Intensity of Attitudes and Behavior." In Robert C. North, Ole R. Holsti, M George Zaninovich, and Dina A. Zinnes, *Content Analysis*, Northwestern University Press.
Zellner, Arnold and Claude Montmarquette (1971). "A Study of Some Aspects of Temporal Aggregation Problems in Econometric Analyses." *The Review of Economics and Statistics*, Vol. 53, No. 4, pp. 335–342.

CONTRIBUTOR

Steve Shellman is a Visiting Assistant Professor in the Department of Government at the College of William & Mary. His research and teaching interests include comparative politics, international relations, and political methodology. His research focuses on conflict processes and the causes and consequences of forced migration. In his dissertation, he posits a theory and constructs a model of dynamic internal conflict-cooperation processes.

APPENDIX: TURN DATA

Recently, scholars have become more interested in sequential modeling techniques (Bakeman and Gottman, 1997; Dixon, 1988; Marlin-Bennet, Rosenblatt, and Wang, 1991; Moore, 2000). To study sequences of state and dissident interactions, Moore (1998, 2000) aggregates events by "moves and turns" as opposed to weeks, months, quarters, and/or years. Marlin-Bennet, Rosenblatt, and Wang (1991, pp. 202–203) introduce an effective distinction between a "turn" and a "move." A move is a single action taken by one actor toward another, while a turn is defined as an uninterrupted sequence of moves by one actor directed toward another. I illustrate the difference using the following sequences of actions:

$$S_1, D_1, S_2, D_2, S_3, D_3, \ldots \qquad \text{(sequence 1: turns)}$$

In sequence 1, the state, S, and the dissidents, D, take "turns" acting and reacting to one another's behavior. However, not every action exhibited by an actor is reacted upon immediately. Time passes between these actions and a group may make two or more "moves" before its opposition makes a move or reacts. Therefore, the sequence looks more like this:

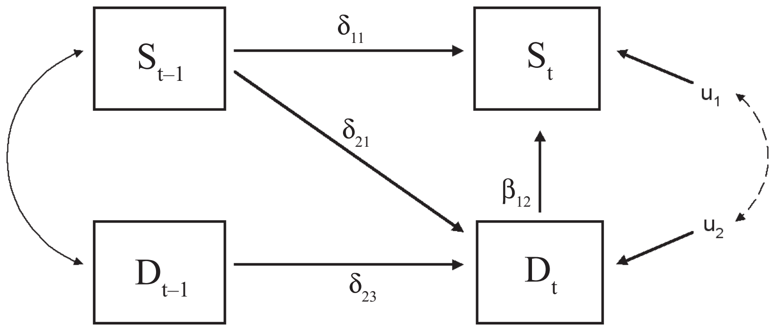
$$s_1, s_2, s_3, d_1, d_2, s_4, d_3, s_5, s_6, d_4, \ldots \qquad \text{(sequence 2: moves)}$$

To analyze behavior in a "turn-taking" fashion, one must produce an ordered sequence of interactions resembling the first sequence. In hypothetical sequence 2, each lower case entry is a "move," whereas the capital letters in sequence 1 represent turns. The first three letters in sequence 2, s_1, s_2, s_3, represent the "turn," S_1 in sequence 1. The next "turn," D_1 identified in sequence 1 is actor D's first and second "move," d_1, d_2 in sequence 2. Actor S's second "turn," S_2 consists of only one "move,"

s_4.

In order to produce the desired sequential "turns" depicted in sequence 1, one aggregates the data by calculating the mean of each sequence of moves.[40] The mean of the sequential moves, s_1, s_2, s_3 now equals turn 1, S_1. For example, we calculate the mean magnitude of actor S's first three moves, in sequence 2, to produce S's first turn. Next, we calculate the mean magnitude of actor D's first two moves, in sequence 2, to produce D's first turn. S's second turn consists of only one move so the magnitude of the move alone represents S's second turn. In effect, series of move interactions is transformed into a series of turn interactions. The data are in sequential order in which the dissidents and the state take turns acting and reacting.

To illustrate how to model sequence 1, we write



or

$$S_t = \delta_{11}S_{t-1} + \beta_{12}D_t + u_1$$

$$D_t = \delta_{23}D_{t-1} + \delta_{21}S_{t-1} + u_2$$

where
S_t = State action taken at time t
D_t = Dissident action taken at time t
S_{t-1} = State action taken at time t-1
D_{t-1}=Dissident action taken at time t-1

Assuming the errors are uncorrelated, we estimate each equation using OLS. If the errors are believed to be correlated, we should employ a three-stage least squares approach or a full information maximum likelihood (FIML) approach.

Copyright of International Interactions is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.