International Studies Review (2007) 9, 437–456

Understanding Data Quality through Reliability: A Comparison of Data Reliability Assessment in Three International Relations Datasets[1]

Steven B. Rothman
Department of Political Science, University of Oregon

Although recent data creation efforts in international relations have begun to focus on issues of reliability and validity more explicitly than previously, current efforts still contain significant problems. This essay focuses on three recent data generation projects that study international relations (the ICOW, ATOP, and River Treaty datasets) and shows the successes and failures of each in assessing reliability when generating data from qualitative evidence. All three datasets attempt to generate reliable data, document the procedures used, and present indications of data reliability. However, their efforts face problems when assessing the reliability of their case selection variables, in the development of reliability indicators, and in the presentation of reliability statistics. In addition to evaluating these recent efforts to generate large-N databases, this essay clarifies the difference between generating data from qualitative and quantitative evidence, explains the importance of reliability when coding qualitative evidence, and provides ways to improve the assessment of the quality of one's data.

Although datasets frequently document the breadth of the cases and variables covered, less attention is paid to assessing the quality of data in terms of reliability and validity. The extent to which data represent the concepts of interest (validity) and the extent to which data are consistent across cases and time (reliability) are important for generating sound conclusions concerning one's findings. Recently, attention has increasingly begun to focus on determining the validity of one's data (Adcock and Collier 2001; Goertz 2005). It is the purpose of this essay to expand this discussion by examining three recent datasets studying international relations and the attention they pay to reliability. This essay will explore the way the three datasets approach the assessment of reliability when compared with ideal criteria for generating such reliability indicators. More spe-

[1] This essay benefited immeasurably from the assistance and support of Ronald Mitchell and from suggestions by Ken Conca, Gary Goertz, Paul Hensel, Brett Asley Leeds, Craig Parsons, Priscilla Southwell, participants in the ''Coding across the Discipline'' (2006) conference directed by Stuart Shulman and the Shambaugh Conference, ''Building Synergies'' (2006), directed by Sara McLaughlin Mitchell. This essay is based in part upon work supported by the National Science Foundation under Grant No. 0318374 entitled ''Analysis of the Effects of Environmental Treaties,'' September 2003-August 2007 (Principal Investigator: Ronald Mitchell). Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author and do not necessarily reflect the views of the National Science Foundation. A previous version of the piece was presented at the meeting of the International Studies Association in Chicago in February 2007.

Published by Blackwell Publishing, 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford OX4 2DQ, UK.

cifically, it examines the extent to which the datasets assess reliability for all variables, generate reliability indicators in a way that makes them unbiased indicators of random measurement error, and present reliability indicators fully for data consumers. The three datasets studied include the Issue Correlates of War (ICOW), the Alliance Treaty Obligations and Provisions (ATOP), and the River Treaty datasets.

The three datasets were chosen because they: (1) make their coding procedures publicly available, (2) have paid some minimal attention to reliability, and (3) vary in the ways that they assess reliability. In order to analyze reliability assessments within datasets, the producers of the dataset must provide coding manuals for review, which were available for the three datasets examined in this essay.[2] Also, all three datasets made some effort to address issues of data error. Since previous critiques of datasets have pointed out that often data creators make little or no effort to address issues of data quality (for example, Vasquez 1987), these datasets represent a second generation of data in international relations. That is, all three have at least made some attempt to assess the quality of their data in terms of reliability. Finally, the three datasets allow us to look at several different techniques for assessing data reliability because they vary in the way that they evaluate reliability and in their presentation of reliability indicators.

The first section of this essay briefly introduces the concept of data reliability as a way of assessing the error in one's data within the context of measurement theory more broadly. In addition, it demonstrates why assessing data reliability is an important aspect of data quality control especially for data generated from qualitative evidence (much of the data used in international relations) and introduces two reliability indicators (Cohen's Kappa and Percent Agreement) with general reliability assessment procedures. The second section proceeds to analyze how the three illustrative datasets have assessed reliability with respect to defining the population of cases, coding procedures, and presentation of reliability indicators. The analysis highlights both the successes and failures of each dataset and describes ways to improve the assessment of the quality of our data.

The Importance of Data Reliability

Before describing the importance of data reliability, it is necessary to define some key terms that will be used throughout this essay. When creating data, we refer to a single data point as a *score*. The term score can be either a numerical or a non-numerical (categorical or nominal) representation of the evidence. Scores do not need to consist of numerical representations of the evidence, but can be qualitative representations using simplified descriptions. *Evidence*, the most basic form of ''raw data,'' consists of the documents or other objects from which scores are generated (see also Mitchell and Rothman 2006). Evidence may consist of items such as historical documents, newspapers, emails and other correspondence, interview transcripts, and even audio–visual media. Evidence is translated into data through the process of *coding*, which generally follows a set of rules included in a coding manual or coding guide. For example, large amounts of evidence are coded regularly within the government in fields such as eRulemaking and intelligence gathering (Wiebe, Wilson and Cardie 2005; Yang and Callan 2005; Kwon, Shulman, and Hovy 2006).

In its simplest form, reliability refers to the ability of repeated coding trials to lead to the same score (Jones 1971:347; Stanley 1971:356; Carmines and Zeller 1979:11–12). In an ideal world, any evidence coded using the same procedure

[2] Some dataset creators, such as Freedom House (2006), do not make their coding procedures publicly available, which limits our ability to analyze their methods.

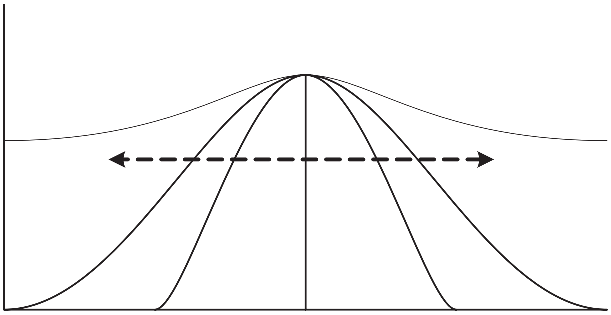

Fig. 1. Decreasing Reliability of a Single Object Coded Multiple Times.

should reveal the same score. Any differences in the resulting scores over repeated trials are attributable to error so long as we can assume that the evidence coded and the coding system have not changed.[3] Intuitively, assuming that nothing changes between coding attempts, we can attribute the difference to error. The important characteristic of this type of error for reliability (as opposed to accuracy[4]) is that this error is random. So, if we were able to take an infinite number of codes of a single particular object for a single particular variable, the result would resemble a normal distribution where the chances of making a positive error (above the true score) and making a negative error (below the true score) are equal.[5] The wider the spread of the resulting scores, the more unreliable the measurement system appears. Note that when this distribution becomes flat (infinitely wide), that is, when there is an equal chance for obtaining any score, the data are perfectly random or unreliable. Figure 1 graphically shows three distributions of measures of a single object in which a wider/flatter distribution represents a more unreliable coding system.

Reliability and Qualitative Evidence

Although assessing data reliability is important for all data, it is especially important for data generated from qualitative evidence. As ideal types, qualitative evidence is distinguishable from quantitative evidence by virtue of the need to interpret such evidence to generate data from it (Ryan 1999; Hovy 2006).[6] Examples of quantitative evidence might involve counting some object where the object is easily distinguishable within the body of evidence. For instance, recording the dates that treaty agreements were signed requires no interpretation of the evidence. The coder must simply read the date from the document and enter that date into some type of database. This transcription essentially involves one simple coding rule. Qualitative evidence, in contrast, requires some level of interpretation and inference and usually several coding rules. Much interesting and worthwhile data for international relations scholars is developed from

[3] If measurement error exists, we want to be able to consider it random and not due to some unknown external variable that has not been controlled. For example, when we measure a physical object, changes in temperature will change the length of some objects as molecules expand and contract. If we are coding text, the external variable could be the degree of tiredness of the coder, for example, resulting in that individual ignoring a word or phrase leading to a different code. For further discussion, see Henry Kyburg (1984).

[4] Accuracy refers to the degree of systematic bias produced by a coding system (Neuendorf 2002).

[5] Non-normal distributions could result when there is a different chance of making errors above and below the mean. There are very few cases of this, but we could imagine a ruler where the numbers below 3 are fuzzy and hard to read, but clear over 3. If the object we are measuring is 3 units long, we are more likely to make larger errors when we under-measure the object than when we over-measure the object in repeated trials.

[6] Although the types of evidence are described here as ideal types, a continuum representing evidence as more or less qualitative and quantitative better represents the true nature of evidence.

qualitative evidence. For example, indicators of countries' state of democracy (Jaggers and Gurr 1995), level of freedom (Freedom House 2006), the content of their treaties (Mitchell 2003; Leeds 2005), and interactions between states (Diehl 2006), all require some level of interpretation of the evidence, some datasets more than others.

Data reliability gains importance as the evidence coded moves from quantitative to qualitative and from directly observable to latent variables.[7] The more interpretation that is required to generate data from evidence, the higher are the chances of introducing additional error.[8] A simple example can illustrate the increased chance of error for more qualitative evidence and latent variables. If data creation requires simply reading a date from one source and entering it into a dataset, it is possible that the coder mistakenly types the wrong number or mistakenly reads the evidence. Given the directly observable information and the quantitative evidence, there are two chances to introduce error. When, however, a researcher codes data from qualitative evidence, several issues make the process more prone to error. Because qualitative evidence does not allow coders to directly observe variables, using such evidence introduces misinterpretation as another possible cause of increased random error. If instead of the date being presented in numerical form, it is presented in words, then the coder must read and understand the textual words to record the data. In addition, if the variable changes from one directly observable to a latent variable (one that must be understood from the text, but is not explicitly present), the chances for misinterpretation increase. If the evidence is textual, the coder may read the evidence correctly and can recite it perfectly, but perhaps his or her understanding of the text is different from another scholar's understanding. The coder may also misunderstand or forget to follow one of the several rules required in coding latent variables because the coding of such variables is generally more complex than coding directly observable variables. As coding qualitative evidence and latent variables usually requires more than a single coding rule and some interpretation, it requires significantly more concentration and memory on the part of the coder, thus increasing the chance for error.

Data in international relations increasingly are derived from qualitative information with many of the variables that are coded requiring interpretation and inference. The assessment of reliability, therefore, is also becoming increasingly more important to doing research in the field. Poor reliability or not assessing data reliability can cause a number of problems for researchers, to which we will turn in the next section.

Problems With Poor Reliability

Low reliability creates several problems for researchers. First, low reliability generates problems when one is trying to find relationships between variables because it means higher rates of error, which change regression coefficients from their true values and reduce confidence in regression results. Second, low reliability creates problems for establishing the validity of the data measures. A dataset is valid when the ''scores meaningfully capture the ideas contained in the corresponding concept'' (Adcock and Collier 2001). Most data are generated for the purpose of studying a particular research question and, thus, the data

[7] The difference between latent variables and observable ones parallels the discussion of qualitative and quantitative evidence except that it refers to the variables rather than the evidence itself. Latent variables are those that are not directly observable and involve interpretation by the coder. Examples of observable variables are recorded dates, number of words, dollars used; examples of latent variables are the degree of democracy of a nation and the implied intent of a particular actor.

[8] This error is in addition to systematic or other error within the evidence itself, which would not necessarily change from the use of quantitative to qualitative evidence or observable to latent variables.

usually reflect the understanding of the concept within that researcher's mind. As data reliability decreases, it becomes increasingly difficult to establish validity because the data become more closely related to random variation than to any systematic concept (Tinsley and Brown 2000:101). The flattest line in Figure 1 shows what a nearly random measurement tool looks like when it is used on the same object repeatedly. Indeed, this nearly random measurement tool indicates little more than randomness; completely random measurement creates a flat line. Poor reliability, therefore, presents researchers with the problems of finding relationships when they exist (increasing the chance of Type I errors) and of establishing that their variables measure the intended concepts.

Benefits of Assessing Reliability

Not knowing the reliability of one's data also creates issues for researchers. In the first place, knowing the reliability of data helps determine an upper limit on correlations with other variables; the higher this upper limit, the more likely scholars are to find statistically significant relationships in their analyses (Slavin 1984:79). If 20% of Variable A's variation is determined by error in the coding system, then by definition only 80% of that variable's variation is systematic or substantive. If we took a second, perfectly reliable variable (Variable B) that we know correlates 100% with Variable A, we would only find an 80% correlation between the two variables because 20% of Variable A is random and cannot by definition correlate with the systematic variation in Variable B. Therefore, knowing the reliability of variables helps us know the upper limits on correlations with other variables, helping us understand the limits on our findings. This knowledge is particularly useful in international relations, where relationships between variables are not generally very high.

As stated above, the possibility of rejecting a relationship when one actually exists between two variables increases with lower levels of reliability. In regressions with error in the variables, the *F*-statistic for the model gets depressed making relationships harder to find (Achen 1983:72). The added knowledge one gets by calculating reliability can help a researcher understand why a relationship between two variables is weak or non-existent when theory may suggest otherwise. Such information also increases our confidence in the resulting inferences and conclusions when reliability is somewhat low but a relationship is found nonetheless. In other words, because low reliability makes it harder to find systematic relationships between variables, scholars can be more confident of a relationship found in noisy, unreliable data.

Knowing reliability also can improve conclusions and inferences by allowing scholars to adjust for error when determining necessary conditions and the magnitude of effects. Reductions in data reliability affect coefficients in important ways. Most texts discuss the reduction in regression coefficients toward zero when data are unreliable; however, these discussions only reflect the average reduction in coefficients when reliability is low in bivariate regressions with a single independent variable (Achen 1983:72). In cases where there are multiple independent variables, one of which has low reliability, the effects on coefficients may be increased or decreased depending on the degree to which the other variables pick up some of the variation that has not been captured by the variable with random error (Achen 1983). In addition, regardless of the size of the error in the independent variable, the magnitude of the bias results from the degree of collinearity. In studying these issues, Christopher Achen (1983) has shown that poor reliability is worse than previously understood because even a small amount of error can have large effects and sometimes in ways not predicted. When researchers have an estimate of data reliability, Frank Baugh (2003) and Bear Braumoeller and Gary Goertz (2000) have provided methods that help them adjust their coefficients. It is well documented that high coding error will affect the magnitude of any effects (King, Keohane, and Verba 1994:158), thus having indications of coding error becomes important in adjusting one's findings. Once the reliabilities of the variables are known, adjusting the coefficients for the random error produces more accurate coefficients in regressions.

Finally, knowing reliability is important because it can improve the quality and clarity of one's coding manual.[9] The development of rules for human coders is often done as part of the research process and in the development of a coding manual or codebook. The coding manual is the tool used for describing the empirical definitions of the concepts or the operationalizations of the variables under study (Pedhazur and Schmelkin 1991:170). Reliability partially measures the degree to which coders understand the rules in the coding manual and can apply these rules to the evidence to generate scores. If reliability indicators are low, researchers can return to the coding manual and clarify points that caused inconsistencies in the coding.[10] This hermeneutic process usually requires several iterations and substantial communication with and among the coders as well as reviewing their codes (Neuendorf 2002). As the coding manual and the concepts become clarified and consistency and reliability among coders increases, it generally is reflected in increased clarity in the coding system. Therefore, users of the data can be confident in their understanding of that particular dataset with a careful reading of the coding manual.

To summarize, high data reliability reduces the likelihood of rejecting a relationship between variables when one exists, reduces biases in the regression coefficients being calculated to examine such relationships, and helps scholars demonstrate that their data represent the intended concepts (that the data are valid). In effect, assessing reliability helps scholars understand why a relationship may be rejected contrary to theory, adjust their regression coefficients for error rates, and improve their coding systems.

Evidence and Coding Procedures

Some further points regarding coding procedures are relevant for generating reliability indicators. In order to determine the reliability of one's data, it is necessary to have multiple scores for each variable-case coded. From measurement theory, we know that if we could generate a very large number of scores for a single variable-case using the same coding tool, we could estimate the distribution of scores (Jones 1971). Fortunately, obtaining two scores for each object over multiple objects creates a distribution of errors, the average of which reflects the random error across the objects. In other words, it is not necessary to generate a large number of scores or measures for a single object if we have two scores for many objects. Whether the means of the error distributions are different or equal, obtaining scores for multiple objects exactly resembles multiple scores from a single distribution.

There are two important considerations when generating multiple scores for a single object: the independence and similarity of the coding processes (Sullivan and Feldman 1979; John and Benet-Martinez 2000). In the first place, coding conditions must be similar to prevent external variables from influencing the scores because reliability is a measure of random error. How can researchers go about creating a most similar coding process for coders? Training coders at the same time and assigning them individual and overlapping sets of cases or objects is one way to achieve arriving at scores for a set of the cases that are indepen-

[9] Although there are relatively few guidelines for creating a coding manual, some descriptions can be found in Kimberly Neuendorf (2002), Abraham Oppenheim (1992), and Charles Smith (2000).

[10] See Daniel Hruschka (2004) for a brief discussion and an example of this process.

dent yet similar (Foster and Cone 1986). Such joint training helps maintain external controls on the two scores by keeping the processes as similar as possible (Holsti 1969:135). In addition, it is important for coders to use exact duplicates of the assigned coding manual as their guide for coding, again so that the processes are similar. Because reliability partially depends on an easily understandable coding manual and coding categories, those that are too complex or contain subject-specific language may interfere with coders' abilities to remain as close as possible in their processes as they introduce their own personal knowledge and interpretations. If coders introduce their own personal knowledge, they are not likely to do so in the same way, which can result in lower reliability. In addition to clarity in coding manuals and categories, prior knowledge of the coding material may generate increased random error. If a coder has some prior knowledge of the coding system, in terms of variables or the objects coded, they may deviate from the coding manual out of confidence in their own understanding rather than applying the rules in the manual. In order to assure that there is no collaboration among coders in the process of coding, they generally do not view each other's codes or discuss the codes. In this way, they are not interfering with or contaminating each other's scores. The only caveat regarding the use of two coders comes when both coders drift from applying the coding rules in the same direction. If both coders shift in the same direction, the resulting data may be biased. Although it is possible for two coders to introduce the same bias into the coding system, the introduction of error in this way is much more likely when only one coder is used.

A second important consideration, hinted at above, is that the coding processes involving the two coders must be independent to avoid the introduction of external factors that may bias the reliability indicator. If the first score partially determines the second score, then there are external elements we need to control that are partially interfering with the second score. It is possible to achieve independence of coding by having the coders work independently from one another when coding the same variables and cases. Consultations, discussed in more detail below, should be kept to a minimum between coders in order to avoid the introduction of additional error into the dataset.

The independence of trials is especially important. Coding processes that are not independent will inflate reliability estimates, whereas introducing more variation (error) by not using a most-similar design reduces reliability estimates. For example, in an extreme case where two human coders code a set of data, but the second coder views the first coder's scores and then codes based on them, the coders are not independent. As higher reliability indicates higher quality, artificially inflating reliability by reducing the independence of coders is appealing but must be avoided. Both maintaining independence among the coders across the coding process and having two coders code for the same categories (a most-similar design) are important for generating accurate reliability indicators. Table 1 summarizes this discussion.

An important caveat on the above discussion involves the use of arbitration or reconciliation in coding processes. When coders consult one another to agree on a score, it reduces the independence among the coders and reduces the validity of the reliability measure, if that measure is taken after the final codes

Table 1. Effects of Varying Assessment Techniques on Reliability Indicators

	Independent assessment	Not independent assessment
Coders use similar coding tools	True reliability	Inflated reliability
Coders use dissimilar coding tools	Decreased reliability	Ambiguous

are produced. In addition, reconciliation coding may include some type of measurement technique not included in the coding manual because the coding involved some arbitrary discussion among coders. Therefore, the coding manual, which should represent all the procedures and rules for developing the data, does not include information on what was said during the arbitration. There are several possible biases that can occur during arbitration, such as the consistent dominance of one coder over another, thus increasing the inaccuracies in the data by generating systematic bias. However, it is important also to note that arbitration can improve the accuracy of coding. Given good group discussion and dynamics between the two coders, it is possible that a discussion of scores could result in a more accurate final code. Although this is a possibility, because biasing the results is also a likely result, researchers are better off to randomly select either coder's final scores when there is a discrepancy between two coders in order to maintain the integrity of the coding system, to avoid introducing coding elements not included in the coding manual, and to generate valid reliability indicators that reflect the final data included in the dataset.

Reliability Indicators

An important distinction among indicators of reliability is that some deal with the covariance among raters and others deal with interrater agreement. Indicators that describe covariance are best used when the type of coding involves assigning values based on ranks rather than on absolute codes. For example, if we asked a panel of scholars to assign scores to countries based on their degree of democracy from one to ten, the scholars are likely to have different averages between them. That is, one scholar may consistently assign higher democracy scores to all countries than another scholar. In this case, however, the ranking of the two scholars might be identical. If one wishes to compare the consistency of the two rankings based on the scores, covariance measures can be used because they measure the degree to which the two scholars vary, whether or not they agree on the exact score they assigned to any particular country. Spearman's *R* statistic is an example of such a covariance indicator.

In most cases in international relations, however, the data themselves are important, not simply the ranking of cases. When the score assigned is important, than interrater agreement indicators are more appropriate. These indicators measure the extent to which two or more raters match exactly on the scores assigned. In the case of the scholars assigning democracy scores described above, when the two scholars assign different scores, the agreement indicator would decrease because the scholars disagreed. This occurs regardless of the rank order of the scores the scholars assign. There are numerous ways to calculate interrater reliabilities; however, the two most commonly used indicators are Cohen's Kappa and Percent Agreement (Neuendorf 2002). Percent agreement simply calculates the number of times coders agree as a percentage of the number of times they could agree. Cohen's Kappa calculates the same statistic but subtracts out the chance agreement. This indicator, thus, is said to calculate the beyond-chance agreement between two coders. Generally, this means that Cohen's Kappa indicators will be lower than percent agreement because the former only illustrates the agreement that occurred beyond what would be expected by chance.

Three Datasets and Reliability

Now that it is clear why data reliability is important for scholars who are involved in building datasets, let us examine three datasets that are currently being used in studying international relations with regard to the attention each pays to reliability and the presentation of reliability indicators. As an important caveat, note that it is not possible to assess each dataset's actual reliability indicators without access to the raw data or recoding the data in some cases. The analysis that follows, however, still shows several areas that could be improved in calculating and presenting data reliability. The three datasets that are examined are the Issue Correlates of War (ICOW) dataset (Hensel and Mitchell 2006), the Alliance Treaty Obligations and Provisions (ATOP) dataset (Leeds 2005), and the River Treaty dataset (Conca, Wu, and Mei 2006). These three datasets were chosen because they all attend to reliability in some way, present their coding procedures explicitly so that they can be examined, and utilize qualitative evidence. The use of qualitative evidence makes reliability more important in each of these datasets and the explicitness of their coding procedures and data allow us to analyze what has been done. This section of the essay will introduce each dataset, describe the data they seek to generate, and delineate how well each of the datasets adheres to the procedures just discussed for generating and presenting reliability indicators. The section that follows then will assess the degree to which all three datasets have adhered to these procedures. In the course of this discussion, the present author will provide some solutions for dealing more generally with the reliability issues that result.

The Issues Correlates of War (ICOW) dataset was designed to generate indicators on ''contentious issues in international politics'' (Hensel and Mitchell 2006). This dataset is intended to show how states resolve their conflicts over particular issues as not all issue conflicts lead to war. In addition, the dataset is constructed to determine which resolutions to these issue conflicts are more likely to succeed in resolving the initial conflict. Thus far, the ICOW project has generated data on territorial claims, maritime claims, and river claims. During the earliest stages of data creation, attention to reliability was generally non-existent, but the authors tried to remedy this situation in a subsequent conference presentation (Hensel 1998). In order to assess the processes for establishing the reliability of their data, we will use both the codebooks of the three datasets created by the ICOW team as well as the subsequent conference paper describing the reliability and validity of their data.

The Alliance Treaty Obligations and Provisions (ATOP) data program collects data on the ''content of military alliance agreements signed by all countries of the world between 1815 and 2003'' (Leeds 2005). The project attempts to understand alliance formation and termination and the effectiveness of formal alliances on changing state behavior (Leeds 2005). The data collected involve aspects of alliances, such as membership, the term, and obligations of member states (Leeds 2005). The codebooks provided by the ATOP principal investigators (PI) are used throughout this essay to evaluate the attention the project paid to reliability.

Whereas ATOP and ICOW are large data collection projects, the River Treaty dataset represents a smaller data collection effort that required less time and fewer monetary resources (which may be more typical of data created by scholars and particularly those engaged in dissertation research) and covers data on river treaties for countries from 1980 to 1992 (Conca et al. 2006). This dataset presents indicators on the attributes of river treaties in order to determine whether states converge on governance over these rivers and whether that convergence influences treaty design. The dataset itself is made available on the Internet and accompanying documentation exists within the Excel workbook and within a published article where the data were first introduced (Conca, Wu, and Mei 2006). The River Treaty dataset compiles various aspects of river treaties including, but not limited to, data on dispute resolution, meeting regularity, and environmental provisions.

Reliability of Case Selection

In almost all data development, there are multiple variables coded for each case. In the ATOP data, for example, a case is identified by a particular alliance treaty and is then coded for various attributes, such as duration, formalization, and obligations. Thus, in an Excel sheet, the cases appear in each row and the various variables in each column. It is necessary for scholars to calculate reliability indicators for each variable coded primarily because consumers of the data often use a limited number of the variables presented (Neuendorf 2002). One variable that is often overlooked is the way in which the cases were selected.

Defining the population of evidence one is going to code is an important part of the data creation process. Without adequately defining the population of evidence, conducting large quantitative analysis may be biased because selection criteria do not fully cover all possible cases (King, Keohane, and Verba 1994). Selecting cases for data collection is especially important because the external generalizability of a particular study stems from proper case selection. For example, quantitative research assessing the effects of international treaties must include the population of all international treaties or select a random subset in order to avoid biasing its results. Such a process involves decisions regarding the nature of the population of cases under study.

Case selection is a form of coding, or categorizing, all the units that might be analyzed into a set that has interest for the researcher. For instance, in the International Environmental Agreement project, Ronald Mitchell (2003) carefully defined and coded treaties according to whether they were (1) international, (2) environmental, and (3) treaties in order to identify a population of international environmental treaties. In this way, he identified the subset of treaties (international environmental) from all the possible treaties available. This type of case selection involves coding, which suggests that it deserves the same attention to reliability and validity as do other variables in one's dataset.

All datasets must deal with case selection in some form. For example, when generating data for the Correlates of War project, investigators had to decide whether a contested state, such as Taiwan, was to be included in the dataset and how to deal with states when they break up or combine (Correlates of War Project 2005).[11] In other words, cases are selected on the basis of some criteria, which we can identify as a variable (a column in our Excel sheet), and therefore it should be treated just like any other variable in the dataset. When identification of cases by some criteria is done with attention to reliability, the PIs who code the cases produce a set of cases that are consistently identified as the cases of interest with a known amount of random error. By considering case identification as part of the coding process, we can determine reliability indicators for this particular variable.

Case Selection in the Three Selected Datasets

The ICOW dataset uses three criteria to select cases to be included in their dataset and implicitly one more criterion in selecting a specific issue. The ICOW dataset focuses on a claim by one state with regard to a particular issue, which is defined as a ''public statement by…an official representative'' (Hensel 2006:4). Other information required to code claims involves the starting and ending dates of a claim, which is determined by the existence or exclusion of a public statement and by searching for evidence that the claim was either settled or abandoned by the state (Hensel 2006:4–5). A third piece of information required to select cases is the determination of whether a claim involves a territory, maritime area, or a river. Given the selection process, there is a chance that two scholars attempting to find cases in the same way may have difficulty finding

[11] Note that because the Correlates of War data generate criteria for whether a particular geographic area is considered a state in their data, the dataset includes case selection in the process of coding.

the same case. Therefore, it is important to treat case selection in the ICOW data the same as coding any other variable in the dataset and assess its reliability.

Case selection within the River Treaty dataset involved the use of treaties from two sources: the Transboundary Freshwater Dispute Database and the FAOLEX legal database (Conca, Wu, and Mei 2006:269). Using these data sources allows case selection to proceed with less concern for error because there is little ambiguity as to whether an agreement appears in one of these sources. However, the authors of the River Treaty dataset add other criteria for case selection that could increase the chances of introducing error by not including agreements that involve ''narrow or isolated matters (navigation, border demarcation, and fishing rights), general agreements in which water played a tangential or trivial role, and agreements unrelated to specific basins'' (Conca, Wu, and Mei 2006:269). The removal of these agreements essentially means that case selection becomes a variable that describes the extent to which the agreement addresses a water basin. The documented procedures for coding do not describe the ways in which, or by what criteria, treaties were eliminated from the dataset used in the final analysis.

The ATOP database comes closest to making case selection a variable, although the reliability of case selection is not indicated. The ATOP cases were selected based on ''written agreements, signed by official representatives of at least two independent sites, that include promises to aid a partner in the event of military conflict, to remain neutral in the event of conflict, to refrain from military conflict with one another, or to consult/cooperate in the event of international crises that create a potential for military conflict'' (Leeds 2005:5). The ATOP dataset has clear documentation on procedures for identifying treaty agreements because many of the existing lists are incomplete or biased for various reasons (Leeds 2005:6–7). At first, the coders collected any agreement that could be considered an alliance treaty, whether it fit the exact definition or not. After all the possible agreements were collected, the PI and the coders decided on whether they fit the exact definition thereby creating two sets of agreements: those that were coded as alliance treaties and those that were not.[12] The PI checked all the cases selected by the coders, but did not do this independently or calculate reliability estimates for the case selection variable. Although the coding by the PI and the coders was not done independently and reliability was not calculated, the ATOP data project has written down its coding procedures explicitly and gathered more cases than necessary in an effort to establish explicit case selection criteria.

Possible Solutions

None of the three datasets described above clearly obtains reliability indicators for the case selection stage of their database development and thus all fail to develop reliability assessments for this important variable in their datasets. There are, however, two basic solutions to coding case selection as a variable that these datasets could have used if a predefined set of broader cases is available.

In both the ATOP and the River Treaty datasets, cases were selected from a larger list of treaties that are available. In both cases, a larger population of cases was available from which to determine what should count as a case for that particular dataset. The simplest way to select cases, therefore, would be a text search of treaty documents.[13] Even when text searches by computers are not appropriate or feasible, it is possible to establish some rules that human coders can use to include a case in the final dataset on which reliability could be assessed.

[12] This discussion regarding the PI's activities was provided via an email conversation with Brett Ashley Leeds in 2006).

[13] For an example of such a process, see Mitchell (2003).

Where there is not a clearly defined larger population of cases available, the PI may have to generate that larger population. The ICOW dataset, for example, has no clearly defined larger population from which it can draw its cases. Instead, the PIs had to generate cases themselves to include in the dataset. The ICOW dataset examines historical texts accessible through libraries to find territorial claims with no limit on the number or types of texts examined.[14] This process makes simple text searches or two-coder reliability checks more difficult, but not impossible. To do so, it would be useful to separate the case selection process from the coding process to make assessing the quality of the case selection process easier. Then, one way to assess reliability would be to limit the sources by selecting several texts for coders and asking two coders to find territorial claims within those sources. This procedure could be done several times using different sources to see how well the coders do at consistently finding the same territorial claims. The resulting indicators will not perfectly assess the reliability of the coding of case selection, but they will give some idea concerning how reliable the territorial claims case selection is and the amount of random error entering into that particular variable.

A second alternative involves asking two coders to find territorial claims in the library using the criteria presented in the coding manual in exactly the same way at different times, with an emphasis on erring on the side of inclusion.[15] Thus, the two coders would be using the exact same procedures to find territorial claims but at different times so that the same resources/books are available in the library. If the coders find the same set of territorial claims, then the process is consistent and reliable. This process could be done asking coders to photocopy texts from sources they believe may involve a territorial claim. By emphasizing that coders should err on the side of inclusion, the number of cases selected will be greater than the number of cases actually to be included in the dataset, but less than the total number of potential cases. After all the materials and evidence are collected, they can be coded for the case selection variable as well as the other variables in the dataset with some attention to reliability.

In general, the process of case selection should be done so that it emphasizes the inclusion of false positives and minimizes false negatives. Because all cases are coded at a later date for the other variables in the dataset, it is better initially to include extra cases where their inclusion in the final population of cases is unclear. That is, if the details of a case make it such that the coder is unclear whether the particular case should be included in the population of cases or not, it is better to include that case from the start than to eliminate it and bias the results. If the questionable cases are included (and noted as such) and later coding of these cases produces some null results on variables of interest, this finding would suggest that these cases may not belong in that particular population. Essentially, this process is what was followed implicitly in the development of the ATOP dataset but without explicit attention to reliability and the possibility of error in case selection.

Reliability Indicators

Reiterating the previous discussion, coders must work with most-similar and independent coding procedures to generate accurate and valid reliability indicators. An example of coding in both a similar and dissimilar process best demonstrates these restrictions; the ATOP dataset provides such an example. Its coding system

[14] This discussion is the result of an email conversation with Paul Hensel in 2006.

[15] The ICOW system for defining cases essentially involves allowing coders to find cases in the library without being limited to particular sets of documents or books (besides the limitation that the sources are found within the library).

was designed so that ''each agreement was coded by at least two coders'' (Leeds 2005:8). The two coding trials, however, were not as similar as possible because the coders used different sets of rules: one coder used a numerical scoring system, while the second coded by answering questions about the agreement in prose (Leeds 2005:8). In this process, there are two sources of variation between each score for a single object, the coders themselves and the coding system they used, either of which could be the source of differences between the scores. In a second instance from the ATOP data, all the agreements were coded twice by the project director using the same numerical system as the original coding (Leeds 2005:8). In this case, the same individual coded the material twice at different points in time. Because we can expect that the text did not change, the only possible source of variation between the two trials is within the human coder. Therefore, in the second case, there is only one source of variation, which makes the two coding trials more similar than in the first case. The more variation that can be controlled between the two coding trials, the more the resulting reliability indicator will reflect random error.

Although the first type of coding in the ATOP database was not done having coders use a most-similar process, the coders were most likely independent.[16] Therefore, we would expect that the two trials did not influence each other. In the second case, however, there is less independence because the same coder coded the data twice, which makes it possible that this individual remembered some of the first codes selected, thus interfering with the second trial. In this type of coding, the greater the time between the coding trials and the greater the number of cases, the less likely the coder is to remember codes in the initial instance while coding the second time. The purpose of the second coding at a later date is to establish some way of assessing reliability when there is a lack of funding or resources for using multiple people to code the same cases.

In effect, the ATOP data assessed reliability using a dissimilar design and two independent coders. This process most likely results in reliability indicators that are lower than the true reliability of the data. If the data were re-evaluated by coders who coded the data independently using a most-similar design, the resulting indicator would probably be a more accurate assessment of the reliability of the data that were generated.

The River Treaty dataset presents an example of data coded by independent coders in a most-similar design. The two coders coded a pilot set of evidence or sample from the population in order to establish a reliability indicator for the entire set. This practice is common when there is a lot of evidence that needs to be coded, but the sample must be large enough to extrapolate information about the entire population (Carney 1972:133–146; Lacy and Riffe 1996; Sawilowsky 2003). Therefore, the River Treaty dataset represents a good example of how reliability indicators should be calculated. The process by which the reliability of the River Treaty data were assessed is most likely going to reflect a true indication of the reliability of the variables coded.

The ICOW data represent an example of a dataset that was developed using an expert coding system. All the data were generated by research assistants who coded specific cases, but without there being multiple coders for each case (Hensel 2006:13–18). This system is referred to as an ''expert'' system because those coding the data were trained experts in the coding process, but their work was not checked by a second coder independently using the same coding system.

Although earlier coding did not assess reliability, the ICOW project attempted to gain some insight into reliability by comparing the initial scores coded by the

[16] It is not clear from the coding manual; but given the coding system described, we can probably assume the coders were independent.

research assistants and the finalized codes, which were reviewed by the project director before being included in the database (Hensel 1998:13). Although such a process provides some indication of reliability, it fails to meet the standard for most-similar coding procedures and independence of the coders because the project director did not code in the same way as the assistants; moreover, the project director viewed each assistant's codes when arriving at the final score. Because neither independent nor most-similar coding processes were used, the effects on reliability assessments are ambiguous. If the project director coded the cases independently without looking at the assistants' codes, reliability indicators would probably decrease. If the project director coded the cases using exactly the same coding process as the research assistants, the reliability indicators would most likely increase. Thus, the assertion that the concordance between the scores of the research assistant coders and the project director's review of such codes is high (Hensel 1998:13) may not be an accurate indication of the reliability of the data. It is unclear whether this assessment is too high or too low because the effects of violating both the rules of independence and most-similar coding creates ambiguous results and will depend on the magnitude of the effects of each violation on the data.

To summarize, establishing good reliability indicators requires multiple scores for the same object. These multiple scores should be generated by independent coders using similar coding systems. There are two ways to generate multiple scores for a single object depending largely on whether the coding involves human coders as research assistants or expert coding by a single scholar. Between the two alternatives, generating scores with multiple coders is optimal because this process creates two scores that are both independent but done on similar evidence. Some services are also available that will, for a fee, create coding manuals and generate data with attention to reliability using multiple independent coders and similar coding systems (for example, the Qualitative Data Analysis Program in Pittsburgh run by Stuart Shulman).

If multiple coders are not available for a project, which is the case for many research enterprises (particularly dissertation projects), a single coder can generate multiple scores for a single object using the test–retest or rate–rerate method. The test–retest method was designed for assessing the reliability or consistency of educational or personality tests over time (Carmines and Zeller 1979:37; Pedhazur and Schmelkin 1991:88). This method, slightly modified, can generate multiple scores when the data do not change over time. First, the researcher codes the data for the concepts desired using a carefully written set of coding rules. Then, a few months later (or after some predefined time), the researcher re-codes the data without referring to the original data coded in the first trial (Tinsley and Brown 2000:102). Using a single coder, however, can introduce bias or non-independent scores. There is a greater chance that a single coder will misapply the same rules twice when generating scores through this method, whereas independent coders will be less likely to code incorrectly in the same way—they are more likely to introduce different biases. Independence is reduced when a single coder codes twice, which can artificially inflate reliability estimates. Independence of the codes is assured by making sure enough time has passed between the trials so that the coder does not code the second set by remembering or referring to the first set (Tinsley and Brown 2000:102). When a single scholar codes the same material twice, it is possible that the scholar consciously or unconsciously applies a code not because of the coding manual, but because the scholar had previously applied that code. That is, the second set of codes may be partially dependent on the first set resulting in inflated reliability estimates—as if two coders examined each other's codes before coding.

Reporting Reliability Indicators

There are several benefits from adequately reporting reliability indicators for data: demonstrating confidence in the data, improving the ability to share data, and reducing the necessity to reinvent coding systems. The ability to control for independent variables, to express a clear quantitative level of confidence in one's results, and to examine the importance of each variable are among the reasons for using regression and other statistical techniques. In a similar way, publishing reliability indicators makes the quality of the data transparent for all users by demonstrating confidence that the data were generated in such a way as to capture the concept in a reliable, consistent fashion.[17] Reporting the scholar's confidence in his or her data provides an important indication of the confidence other scholars can place in the fact that the data reflect the coding rules and the concepts the project was intended to code. Presenting indicators of reliability allows other scholars to evaluate the quality of the data used.

In addition, other scholars can use reliability indicators to distinguish between two datasets, when other aspects of the data are equivalent. Given two datasets, where some of the data overlap, the dataset with higher reliability estimations is preferable because the data are likely to introduce less error into one's research. For example, the ICOW project demonstrates convergent validity (see Campbell and Fiske 1959; Carmines and Zeller 1979) by looking at several datasets that coded territorial claims and comparing them with the data generated in the ICOW project. The results show that the different datasets found similar occurrences of interstate disputes during the overlapping time periods covered in each dataset (Hensel 1998:20). If all the datasets examined had published reliability indicators, scholars could have chosen to use that set with the higher reliability indicators. In other words, if two datasets have a similar number of cases coded for similar variables, what makes one better than another? The answer could lie in the dataset that has the higher reliability indicators. Therefore, reporting indicators of reliability can assist scholars in choosing among datasets.

The second advantage of publishing reliability indicators involves the ability to share data and reproduce results using the data. Replication of research has gotten considerable attention as ''the most common and scientifically productive method of building on existing research'' (King 1995:445). It is a fundamental part of the scientific method and vital for a progressive research program. Replication of results sometimes consists of coding data in the same way that was done in the original analysis (King, Keohane, and Verba 1994:26–27). For a researcher who desires to replicate an entire study, from data creation through the interpretation of results, knowing the reliability indicators becomes important.[18] Having reliability indicators available allows the author to compare the reliability from the data in the replication effort to the reliability in the original study. If the second researcher's reliabilities are much lower or higher, it may explain differences in any results that are found. In addition, the availability of reliability indicators can inspire future researchers to develop better coding systems and ultimately better data. If one researcher's reliabilities are 70%, for example, a new coding system might create very similar data with less error, for instance, 90% reliable, allowing researchers to find stronger relationships in their analyses. Therefore, indicating reliabilities in published data allows for the accumulation of data and improves replication.

[17] In addition, just like regression confidence levels, reliability indicators can be expressed with confidence intervals (see Fan and Thompson 2003 for an in-depth explanation).

[18] Replicating some types of data collection may be impossible when they involve interviews or informal surveys, however, coding such data from notes can be replicated if the notes are available.

A third advantage of publishing reliability indicators is to be able to reduce the necessity of reinventing coding systems and the ability to develop a set of generally used concepts, greatly reducing the work of future scholars conducting research in similar areas. Although there is some controversy regarding the generality of reliability indicators, knowing that a single coding technique is reliable across studies and types of data suggests the general reliability of that specific coding technique. Reliability indicators show the amount of random error in particular data, coded by particular coders, at a particular time, for a set of particular evidence (Tinsley and Brown 2000:96; Brennan 2001:301; Sawilowsky 2003). However, when the same coding technique is used by multiple scholars to code similar types of evidence at different times by different coders, having multiple indicators of reliability can help to establish that this particular coding system produces data at a certain error rate. In other words, scholars can conduct meta-analysis on data developed using that particular coding technique (Fan and Thompson 2003). This does not mean that we should assume that the ATOP coding manual, for example, will be useful in coding all types of human behavior, that would clearly be outside the applicability of the ATOP coding system. If several scholars, however, used the ATOP manual to code data from similar evidence with different sets of coders that resulted in similar error rates, we could begin to say that the ATOP coding system and the concepts underlying it could be used by other scholars coding similar variables without worry of a high degree of random error. In this view, reliability indicators do not exist as a trait of the data, but exist as a trait of the coding system used to develop the data. When multiple scholars using the same coding system achieve high reliability, it is possible to show that the coding system is a reliable one rather than attributing reliability to each dataset individually. In medicine and educational testing, for example, there are several standardized coding systems or tests.[19] A necessary condition for determining the reliability of a particular coding system is that scholars who generate data using that system make their coding manual and reliability indicators publicly available. It is only after knowing the reliability indicators for a coding system across a number of studies and conducting meta-analysis on those studies that it becomes possible to talk about generalizable or standard coding techniques.

Presentation in the Three Datasets

Of the three datasets studied here, the River Treaty and the ATOP datasets have explicitly assessed reliability in a way that allows them to publish their reliability indicators. In addition, the River Treaty dataset is the only one of the three that has produced multiple codes using independent coders and the most-similar coding technique, as is generally considered the best practice in generating reliability indicators. The coders developing the ATOP data generated multiple codes, as was indicated earlier, but the project failed to present the reliability indicators for each variable coded in the codebook (Leeds 2005:8). The lack of quantitative reliability indicators makes most of the above reasons for producing reliability indicators irrelevant. However, as the data were created using multiple coders, it seems feasible that the project director could return to the coded data sheets and generate intercoder agreement statistics.

The River Treaty dataset does present reliability indicators for all the variables used and does so within its published results. These PIs (Conca, Wu, and Mei 2006:283) also followed the general practice of not using any variable that fell below a specified level of reliability, in this case 80%. Although the authors of

[19] Consider, for example, some of the standard medical rating scales such as that for blood pressure and intelligence scales such as the Wechsler (Nunnally 1959:191).

the River Treaty dataset provide reliability indicators for each variable used, they fail to describe what method was used in calculating these reliability indicators. Given the large number of possible reliability indicators available to scholars, such as Krippendorff's Alpha, Percent Agreement, and Cohen's Kappa, it is important for authors to be explicit about which of the indicators is being used. Each of the available indicators has its own idiosyncratic characteristics, such as systematically inflating or reducing reliability estimates (see Cronbach 1951; Cohen 1960; Smith, Herrera, and Herrera 1990; Shrout 1998; Lombard and Snyder-Duch 2002; Cronbach and Shavelson 2004).

Solutions

In general, the solution to presenting reliability indicators is to do so explicitly for each variable coded in one's dataset including in the case selection stage. Each variable needs its own reliability indicator because only certain variables may be relevant to other scholars. If reliability indicators are attributed to an entire dataset and not each variable in that dataset, then users are given little information about whether a particular variable is above or below the average across all the variables. Such reporting is generally considered a misrepresentation of the true reliability of the data (Neuendorf 2002:142).

Conclusion

This essay has presented information on establishing the reliability of data generated from qualitative evidence through an examination of three recent data creation efforts—ICOW, ATOP, and the River Treaty dataset. All three datasets are insufficiently attentive to a particular aspect of reliability: in case selection, in the development of reliability indicators, or in the presentation of reliability indicators. In particular, all three datasets fail to assess reliability for their case selection variables. The ATOP data additionally reliability in a way that may bias the indicators developed. The ICOW data develop only rough, qualitative, estimates of reliability in a manner that may also bias the indicators. The River Treaty dataset does not specify which indicator of reliability it uses in its calculations. When generating data, it is important to consider reliability indicators at the case selection stage because case selection is a form of coding for the population of evidence. In order to generate reliability indicators for one's data, it is important to begin the process of coding by using multiple coders that are using independent and most-similar coding systems. Finally, without presenting the reliability of each variable and describing what types of statistics were used to generate the indicators, users of the data cannot adequately understand the error-rate, cannot compare replicated data collections, and cannot compare similar datasets on issues of reliability.

This essay has not claimed that the error in these datasets is abnormally high or low other than what the researchers have provided as evidence of their own error rates. The author cannot make such claims without knowing much more about the attention paid to reliability during the development of the datasets. The emphasis here, instead, has been on using the three datasets to illustrate the importance of assessing reliability during the data creation process and the fact that as consumers of data we cannot know the error rate in particular data unless the PIs assess the quality of their own data and present such indicators publicly.

One last caveat is relevant regarding the differences in complexity among the three datasets evaluated in this essay. Datasets vary based on the breadth and scope of their coverage (for example, in terms of the number of years or countries covered), the complexity of the variables that are assessed (latent variables versus observable variables), and the number of variables measured. When a dataset involves a large number of latent variables and is broad in scope, a trade-off is often made regarding the amount of attention paid to the quality of the data. That is, large and complex datasets are often imprecise (at a lower level of resolution) and the developers spend less time attending to issues such as reliability because the time that is required to code the data itself is usually extensive. In terms of the data projects examined in this essay, all three contain some latent variables, but the River Treaty dataset contains fewer years and variables than the ATOP and ICOW datasets. Therefore, we might expect that the authors of the ICOW and ATOP datasets spent less time addressing issues of reliability than did those developing the River Treaty dataset simply because these datasets are much larger and more complex. In addition, much of what is presented above represents the ideal presentation of data and attention to reliability. Even if assessments of data quality cannot be done for all cases, sampling cases can provide some important insights. Finally, this essay is not arguing simply that assessing reliability will make for higher quality data (although this argument might be true in part for the development of coding manuals and systems). Instead, the emphasis is on the contention that unless there is some reliability assessment in our data development, the consumers of the resulting datasets cannot possibly know the degree of error in the data nor compare error rates with alternative datasets. Therefore, the trade-off between the quality of the dataset and its complexity is misleading in this context. This trade-off is generally made when the database creator decides on the scope and depth of the project, knowing that the quality of the data will decrease with greater complexity. But even having made such a decision, the argument being made here is that it is still important to measure the quality of the data that is created.

The purpose of this essay was not to point out problems in currently used datasets in international relations, but to point to ways we can increase the value of data created in the future. There are many trade-offs when scholars generate data, but assessing data quality does not need to be traded for greater scope or complexity. At a minimum, scholars can provide qualitative indications of reliability based on a small sample of data coded by multiple scholars. In addition, as the complexity of the data development increases (more sources/evidence or greater numbers of latent variables), attention to data quality should also increase because the likelihood of increasing the error rates also rises with complexity. As more scholars move to analyze qualitative evidence, such as texts and discourse, in quantitative ways, developing reliable data will become increasingly important. Data developed in the ways described above will produce higher quality data and more robust and confident findings in the longer term.

References

Achen, Christopher H. (1983) Toward Theories of Data: The State of Political Methodology. In *Political Science: State of the Discipline*, edited by Ada W. Finifter. Washington, DC: American Political Science Association.

Adcock, Robert, and David Collier. (2001) Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95(3):529–546.

Baugh, Frank. (2003) Correcting Effect Sizes for Score Reliability. In *Score Reliability: Contemporary Thinking on Reliability Issues,* edited by Bruce Thompson. Thousand Oaks, CA: Sage Publications.

Braumoeller, Bear F., and Gary Goertz. (2000) The Methodology of Necessary Conditions. *American Journal of Political Science* 44(4):844–858.

Brennan, Robert L. (2001) An Essay on the History and Future of Reliability from the Perspective of Replications. *Journal of Education Measurement* 38(4):295–317.

Campbell, Donald T., and Donald W. Fiske. (1959) Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin* 56(2):81–105.

Carmines, Edward G., and Richard A. Zeller. (1979) *Reliability and Validity Assessment.* Beverly Hills, CA: Sage Publications.

Carney, Thomas F. (1972) *Content Analysis: A Technique for Systematic Inference from Communications.* London: Batsford.

Cohen, Jacob. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1):37–46.

Conca, Ken, Fengshi Wu, and Ciqi Mei. (2006) Global Regime Formation or Complex Institution Building? The Principled Content of International River Agreements *International Studies Quarterly* 50:263–285.

Correlates of War Project. (2005) State System Membership List (v 2004.1). Available at http://www.correlatesofwar.org/.

Cronbach, Lee J. (1951) Coefficient Alpha and the Internal Structure of Tests. *Psychometrica* 16:297–334.

Cronbach, Lee J., and Richard J. Shavelson. (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement* 64(3):391–418.

Diehl, Paul F. (2006) Correlates of War. Available at http://www.correlatesofwar.org/.

Fan, Xitao, and Bruce Thompson. (2003) Confidence Intervals about Score Reliability Coefficients. In *Score Reliability: Contemporary Thinking on Reliability Issues,* edited by Bruce Thompson. Thousand Oaks, CA: Sage Publications.

Foster, Sharon L., and John D. Cone. (1986) Design and Use of Direct Observation. In *Handbook of Behavioral Assessment,* edited by A. R. Ciminero, K. S. Calhoun and H. E. Adams. New York: Wiley.

Freedom House. (2006) *Freedom in the World.* Washington, DC: Freedom House.

Goertz, Gary. (2005) *Social Science Concepts: A User's Guide.* Princeton: Princeton University Press.

Hensel, Paul R. (1998) Reliability and Validity Issues in the Issue Correlates of War (ICOW) Project. Paper presented at the annual meeting of the International Studies Association, Minneapolis.

Hensel, Paul R. (2006) General Codebook: Issue Correlates of War (ICOW) Project. Available at http://www.garnet.acns.fsu.edu/~phensel/icow.html.

Hensel, Paul R., and Sara McLaughlin Mitchell. (2006) The Issue Correlates of War (ICOW) Project. Available at http://garnet.acns.fsu.edu/~phensel/icow.html.

Holsti, Ole R. (1969) *Content Analysis for the Social Sciences and Humanities.* Reading, MA: Addison-Wesley.

Hovy, Eduard. (2006) Toward a ''Science'' of Annotation. Coding across the Disciplines. A Project Based Workshop on Manual Text Annotation Techniques, Pittsburgh.

Hruschka, Daniel J., Deborah Schwartz, Daphne Cobb St. John, Erin Picone-DeCaro, Richard A. Jenkins, and James W. Carey. (2004) Reliability in Coding Open-Minded Data: Lessons Learned from HIV Behavioral Research. *Field Methods* 16(3):307–331.

Jaggers, Keith, and Ted Robert Gurr. (1995) Tracking Democracy's Third Wave with the Polity III Data. *Journal of Peace Research* 32(4):469–482.

John, Oliver P., and Victoria Benet-Martinez. (2000) Measurement: Reliability, Construct Validation, and Scale Construction. In *Handbook of Research Methods in Social and Personality Psychology,* edited by Harry T. Reis and Charles M. Judd. New York: Cambridge University Press.

Jones, Lyle V. (1971) The Nature of Measurement. In *Educational Measurement,* edited by Robert L. Thorndike. Washington, DC: American Council on Education.

King, Gary. (1995) Replication, Replication. *PS: Political Science and Politics* 28(3):443–449.

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

Kwon, Namhee, Stuart Shulman, and Eduard Hovy. (2006) Multidimensional Text Analysis for eRulemaking. Paper presented at the 7th National Conference on Digital Government Research, San Diego.

Kyburg, Henry E. (1984) *Theory and Measurement.* Cambridge: Cambridge University Press.

Lacy, Stephen, and Daniel Riffe. (1996) Sampling Error and Selecting Intercoder Reliability Samples for Nominal Content Categories. *Journalism and Mass Communication Quarterly* 73(4):963–973.

Leeds, Brett Ashley. (2005) Alliance Treaty Obligations and Provisions (ATOP) Codebook. Available at http://www.atop.rice.edu/.

Lombard, Mathew, and Jennifer Snyder-Duch. (2002) Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research* 28(4):587–604.

Mitchell, Ronald B. (2003) International Environmental Agreements: A Survey of Their Features, Formation, and Effects. *Annual Review of Environmental Resources* 28:429–461.

Mitchell, Ronald B., and Steven B. Rothman (2006) Creating Large-N Datasets from Qualitative Information: Lessons from International Environmental Agreements. Paper presented at the annual meeting of the American Political Science Association, Philadelphia.

Neuendorf, Kimberly A. (2002) *The Content Analysis Guidebook.* Thousand Oaks, CA: Sage Publications.

Nunnally, Jum C., Jr. (1959) *Tests and Measurements: Assessment and Prediction.* New York: McGraw-Hill.

Oppenheim, Abraham N. (1992) *Questionnaire Design, Interviewing, and Attitude Measurement.* London: Pinter.

Pedhazur, Elazar J., and Liora P. Schmelkin. (1991) *Measurement, Design, and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum.

Ryan, Gery W. (1999) Measuring the Typicality of Text: Using Multiple Coders for More than Just Reliability and Validity Checks. *Human Organization* 58(3):313–322.

Sawilowsky, Shlomo S. (2003) Reliability as Psychometrics Versus Datametrics. In *Score Reliability: Contemporary Thinking on Reliability Issues*, edited by Bruce Thompson. Thousand Oaks, CA: Sage Publications.

Shrout, Patrick E. (1998) Measurement Reliability and Agreement in Psychiatry. *Statistical Methods in Medical Research* 7:301–317.

Slavin, Robert E. (1984) *Research Methods in Education: A Practical Guide.* Englewood Cliffs, NJ: Prentice-Hall.

Smith, Charles P. (2000) *Content Analysis and Narrative Analysis.* New York: Cambridge University Press.

Smith, Eric R. A. N., Richard Herrera, and Cheryl L. Herrera. (1990) The Measurement Characteristics of Congressional Roll-Call Indexes. *Legislative Studies Quarterly* 15(2):283–295.

Stanley, Julian C. (1971) Reliability. In *Educational Measurement*, edited by Robert L. Thorndike. Washington, DC: American Council on Education.

Sullivan, John L., and Stanley Feldman. (1979) *Multiple Indicators: An Introduction.* Beverly Hills: Sage Publications.

Tinsley, Howard E. A., and Steven D. Brown. (2000) *Handbook of Applied Multivariate Statistics and Mathematical Modeling.* San Diego: Academic Press.

Vasquez, John A. (1987) The Steps to War: Toward a Scientific Explanation of Correlates of War Findings. *World Politics* 40(1):108–145.

Wiebe, Janice, Theresa Wilson, and Claire Cardie. (2005) Annotating Expressions of Opinions and Emotions in Language. *Computers and the Humanities* 39(2–3):165–210.

Yang, Hui, and Jamie Callan. (2005) Near-Duplicate Detection for eRulemaking. National Conference on Digital Government Research, Atlanta, ACM International Conference Proceeding Series.