Conflict Consortium Standards & Best Practices for Observational Data*

Christian Davenport† and Will H. Moore‡

http://ConflictConsortium.com

April 7, 2015

*This document grew out of two conferences the authors convened to collect input from scholars in the field. Benjamin Appel, Mark R. Beissinger, Johanne Birnir, Kanisha Bond, Jessica Braithwaite, Dara Kay Cohen, Marianne Dahl, Laura J. Dugan, Kristine Eck, Scott Edwards, Chris Fariss, Larissa Fast, Omar García-Ponce, Anna Getmansky, Anita Gohdes, Tamy Guberek, Peter Haschke, Courtney Hillebrecht, Linda Camp Keith, Yonatan Lupu, Aila Matanock, Halvard Nygard, Todd Sandler, Philip A. Schrodt, Zachary Steinert-Threkheld, Christopher Sullivan, Ralph Sundberg, Jakana Thomas, Ernesto Verdeja, Reed Wood, Thomas Zeitzoff and Yuri Zhukov were kind enough to participate. Scott Bennett, Michael Findley, Jacob Shapiro and Nicholas Weller have also provided useful comments on a draft. The Consortium has received support for this project from the **National Science Foundation (SES-1144064)**.

†Professor, Department of Political Science, University of Michigan. email: christiandavenport@mac.com

‡Professor, Department of Political Science, Florida State University; after August 2015, School of Politics & Global Studies, Arizona State University. email: WHMooreASU@gmail.com.

Contents

1 What is this Document?

This document establishes standards and practices that researchers who collect observational data on political conflict and violence,[1] as well as those who referee research (proposals as well as manuscripts being considered for publication) that use such data, should strive to meet.[2]

We undertook this project because we have observed, over the course of our careers, that when researchers develop projects few appear to avail themselves of the details and nuance of the data collection efforts that underpin much of the work that they reference (to say nothing of work that is not cited in proposals). We decided that the absence of a single source that collected the "best practices" and set standards for data collection in the study of violent political conflict was a considerable hindrance to data development, analysis and ultimately cumulation.

1.1 How should the document be read?

Essentially, we endeavored to create something that could be either read in its entirety and provide some general guidance, but not to provide too much detail as each subject is worthy of an article in an of itself. The idea would be to read this document before and during the creation of a database in order to get an idea of what should generally be considered. Alternatively, the document could be read while a database project was being considered to assist the evaluator with thinking about what problems normally arise and should be considered.

To assist with this process, at the beginning of the various sections, we boldface the most

[1] Data are also generated in experiments, and while many of the issues discussed in this document are pertinent to such data generation, this document does not address and of the sundry issues that arise in the context, but do not arise in the context of creating observational data.

[2] Readers might also be interested in two reports produced by the "Data Access and Research Transparency (DA-RT) in Quantitative and Qualitative Research Traditions" working group of the American Political Science Association (`http://www.apsanet.org/content_86135.cfm`).

important points so that at a glance one can understand what the takeaway point is from the relevant section. The remaining discussion elaborates on the takeaway point and provides some clarification regarding how we arrived at the point made.

1.2 Disagreement about Standards & Practices

While this document lays out standards and practices for collecting conflict data, we are well aware that standards and practices are dynamic and evolve. Indeed, we hope to facilitate a conversation so that we might improve our craft. Acknowledging this, we have sought, where feasible, to articulate a vision, but there are certain topics where we feel that there is inadequate agreement about the best standards and practices. What to do?

Rather than wring our hands about this situation, we will seek out scholars who can articulate the rationale behind a given standard and practice, and ask them to contribute posts on blog space we plan to create at the Conflict Consortium (CC) site (`www.conflictconsoritum.com`). We ask members of the community to comment, and add their own posts. As convergence emerges, we hope CC members will step forward and revise the standards and practices accordingly.[3] Indeed, the document you are reading is not, and does not attempt to establish, the standards and practices which will thereafter set the metric against which all work in the area should be evaluated. Rather, we view it as a dynamic document that will be revised periodically: we acknowledge that we are dealing with something of a moving target and intend this document and the site's blog space as tools to facilitate discussion about how the community can better hit the target as it moves.

Similarly, we invite community members to post and reblog to the forthcoming blog space at `www.conflictconsoritum.com` their case against a standard and practice contained

[3]We are forming a Board of Directors and envision that future revisions to this document could be undertaken by an ad hoc committee of members, then submitted to the Board for approval. Other options exist. For example, a Consortium member might undertake to creating a wiki version of the document, which would allow community members to revise it organically. Our goal is to create the forum where the community can undertake such endeavors.

herein, and advocate for an alternative. As appropriate, we expect that CC members will form an ad hoc committee to revise this document.

1.3 Is All Conflict Data Event Data?

The dichotomy between "event" and "standard" data is false.

The disagreement that can work against the development of standards and best practices are well illustrated by a debate that is summarized in chapters 4 and 5 of Landman & Calvarho's 2009 book *Measuring Human Rights*. They discuss, and weigh in on, a tension between projects that code "who did what to whom, when and where" (generally known as "events data") and projects that select a fixed spatial-temporal unit of observation (frequently the country-year) and assign values to each observation along an ordinal scale (generally known as "standard based data"). Which of these is better? More directly, does the "sub-national, sub-annual" (aka disaggregating conflict) movement threaten standard based data collection, and if so, is that a problem?

We believe the event v standard based dichotomy is false. As we discuss below (in the Units of Observation section), both space (e.g., village, city, longitude-latitude) and time (e.g., hours, days, weeks, months) are continuous domains that can be carved up into discrete units, but nevertheless remain in principle continuous. Our failure to recognize this fundamental fact leads us to fail to imagine alternatives that have always been available to us, and practice thus contributes to widespread perception of a false dichotomy. This, then, is the good news: ***we need not select between these two***![4] Indeed, the vast majority

[4]We hasten to add that the more fine grained in space and time one wishes to work, the more missingness one introduces. Thus, in practice, one will rarely find it possible to code complete cases for the most fine-grained spatial and temporal units reported within any given source: the granularity at which one can code will often vary within any specific source. We are advocating the coding of each piece of information at the most fine grained level that piece of information permits, thus producing data that can be aggregated, using computers, over units the user selects. The user will quickly realize that the missingness declines as she selects increasingly coarse temporal and spatial units (e.g., year and country). Techniques for addressing missingness are proliferating, and the standard and practice we advocate moves choices about handling missing data and bias from the data collection team to the research analyst. We believe that is a more

of what is called "standard based" data is some sort of systematic aggregation of something that could also be represented as some type of events data.

Let us take a well known standard based data set to illustrate, the Political Terror Scale. This project uses the country-year as the unit of observation and coders perform content analysis of US State Department and Amnesty International reports to assign each country-year a value along a five point ordinal scale. The value indicates the extent to which the government of that country violated the rights of the people within its borders. In the language generally used to describe events data, this project codes the aggregated (or average, typical, modal) annual (when) rights-violating activity (what) of the government (who) against the population (whom). It is standard based because the "what" is coded against a standard (international human rights law). What tends to throw people is that it is aggregated annually not by adding up all events, but instead by conducting a content analysis of a natural language summary of some of those events as identified by the sources employed. We suspect that an implicit focus on the imagined "population" of events contributes to the confusion: it is readily apparent that the natural language summary is not a total or mean of the population of events. Yet folks fail to recognize that an events data base such as ICEWS or SPEED also is not the population of events. It is, instead, a record of "news worthy" events.[5] Seen from this vantage the distinction between events data and standard based data is their spatial-temporal units of aggregation: the former (events) are fine grained on both dimensions while the latter (standards) are more coarse on both dimensions. The granularity that best serves a given research project cannot be determined independent of that project. That said, to minimize effort, we should collectively endeavor to develop standards and best practices that make it possible to code a given source as few times as possible.

appropriate location for such decision making.

[5]This played out on social media in 2014 in response to a post by Mona Chalabi at `fivethirtyeight.com` using the GDELT data: How to GDELT Properly and Mapping Kidnappings in Nigeria (Updated).

2 Planning your Data Collection

2.1 What Concepts to Code

A tension exists between creativity v standardization, and researchers should consider the tradeoff between conceptual innovation and broad engagement of existing research. When innovating, consider whether compatability with past efforts is feasible.[6]

When beginning a data collection effort one is confronted with where to place one's conceptual effort relative to existing data collections.[7] In doing so, it is generally useful to consider the variety of conceptual definitions in the literature relative to the research questions that the various authors considered. It is far from obvious that a single conceptual definition of, say, *counter-insurgency*, *dissent*, *ethnic conflict*, or *state terror* is useful for all of the research questions the community asks about any given phenomenon. It is obvious that some definition should be provided.

To the extent the researcher finds herself dissatisfied with the status quo conceptual definitions she is motivated to propose a novel definition. Alternatively, she may find herself satisfied with existing conceptual definitions, but dissatisfied with existing operational definitions. The latter situation is less complex than the former, but we remind researchers to consider the extent to which their choices about conceptual and operational definitions will impact the extent to which other researchers will engage the data.

Measures of widely shared definitions will engage a wider community, thereby supporting cumulation of knowledge. That said, advances often occur when someone introduces a new coneptual (or operational) definition, and supplants the status quo. How should one

[6] As noted above, by existing research we do not mean work published in the past decade, or included on relevant graduate seminar syllabi, but relevant work published at any time in the past.

[7] It is important to consciously distinguish concepts from operational indicators of those concepts. Failure to do so makes it impossible, for example, to assess the validity of a given indicator. This sub-section addresses concepts.

weigh the benefits of cumulation against the benefits of innovation? We are unable to offer guidelines on how to assess the likely benefits and costs (to the individual researcher or the community) of choosing between status quo conceptual and operational definitions versus innovating one or both when undertaking data collection. We simply urge researchers to set aside some time to think about, and discuss with colleagues, these issues.

That said, we encourage researchers who adopt an innovative operational approach to consider whether they might be able to also code the source material to produce an indicator compatible with existing indicators.[8] In some cases, the use of finely grained data (which we discuss below) can be used to permit conceptual and operational flexibility, including the potential for producing both innovative and standard measures from a given set of source material. Why care about comparability? There are numerous reasons: accumulation of knowledge, replication and assistance to the broader community who may wish to understand how the newer effort relates to others.

We should also observe that while much measurement is done in a reasonably direct fashion, innovative measurement often involves thinking of creative use of proxies that are readily measurable to get at concepts with values that are not readily recorded, making use of instruments, and so on. Though this document focuses attention on what one might call direct measurement, we would be remiss not to observe the value of innovative ideas about proxy measures, etc.

2.2 Population and Sample

Researchers should state both what is their population, and whether their data represent a census or a sample of that population.

Our community tends to invest precious little consideration in whether a given data

[8] As an example, compare the Ill Treatment and Torture project's `Level of Torture` variable, which is compatible with Hathaway (2002), with that project's innovative measures of states' use of ill treatment and torture (Conrad, Haglund and Moore, 2013, 2014).

collection represents the population or a sample. A population is the full set of cases (aka a census). A sample is a subset of the population that is missing some of the cases from that population. One would think that distinguishing populations from samples is elementary, but it turns out to be more complex than one would initially suspect. For example, the CIRI data (Cingranelli and Richards, 2010) is a content analytic condification of both of the following:

1. the population of reported rights abuses contained within the annual reports of Amnesty International and the United States Department of State;

2. a convenience sample of the population of "naming and shaming" activities by INGOs and countries.

Reasonable arguments can be made to defend both of these views, and as a community we need to invest more effort considering, and explicitly defining, whether a given data collection project, or research effort using given data, are treating the data as a population or a sample.

Consider a second example, the well known case of an events dataset built upon news sources. Any such database *is* the population of news reports by the news organizations whose reports were coded, but it *is also* a convenience sample of the events of interest (specifically, those events that the news source reported) from the population of all such events that took place within the temporal–spatial domain of the population. These are important distinctions, and we discuss both in more detail below.

The key point, however, is that researchers who collect conflict data should state explicitly what is their population, and then describe the sampling strategy employed, thereby making clear what relationship exists between the sample and the population of interest. Using the terminology from survey research, those who collect data should describe the sampling frame and then the technique used to sample from that frame (Cochran, 1953).[9]

[9]For a good online source, see the Research Methods Knowledge Base's Sampling entry (Trochim, 2000).

Put differently, researchers who collect conflict data should state explicitly whether their data can be conceived of as a population,[10] a sample, or both, and the circumstances under which they encourage users to treat their data as one versus the other. That information should be prominent in the User Guide made available to those who use the data after it is released/published.

2.2.1 Coding Populations

To be clear, it is unusual for a conflict dataset to explicitly state whether it is coding a population. Yet, many such efforts exist. For example, many of the Correlates of War datasets code every example of the phenomenon they are measuring that occurs among the population of interstate members (as that project defines them). Such an effort is a population (or census) approach to data collection, and these are perfectly viable.

One common type of population conflict dataset that exists is rarely consumed by the research community as population data: content analytic databases that code the full population of a certain type of document, such as Amnesty International's Annual Reports from 1980-2000, or all of the articles published by Agence France Press in 2010. The ICEWS and SPEED projects, which code a population of news reports, are two of many such examples (Boschee et al., 2015; Nardulli, Althaus and Hayes, 2015). This example helps illustrate that one can think of a single dataset as both a population and a sample. The ICEWS and SPEED projects, and other events data projects like them, are both a population of news reporting by a given set of news organizations, and at the same time a convenience (systematic) sample of all news reports, as well as a convenience systematic (sample) of all interstate interactions. A researcher who wanted to study how the news agencies coded by ICEWS or SPEED reported the news, such as Baum and Zhukov (2015), would treat those data as a population.

[10] To the extent it can be treated as a population it is critical to clearly demarcate that population.

The most likely reason that these sorts of efforts are not typically understood to be populations is because researchers tend to want to use them *as if* they were measures of the extent to which governments respect their international treaty obligations with respect to human rights (rather than a record of Amnesty International's allegations of violations) or as a record of the actual events of a given type that occurred in the world (rather than a record of what Agence France Press published about such events). This distinction is not trivial. Indeed, conceived as a population, these are perfectly valid data that, depending upon coding schemes, might be highly reliable. Yet understood as samples, they are convenience samples that are non-random, and we discuss below the challenges this raises.

2.2.2 Coding Convenience Samples

A sample is a subset of a population and can be broadly categorized into two types: random and systematic. Simple random samples, of approrpiate size, have the desirable property that they are representative of the population along all dimensions one might conceive. Systematic samples are not representative (by definition), though any given systematic sample might be. More specifically, a systematic sample will be biased, producing central tendency values that are too high or too low.

A convenience sample is a systematic sample that was collected for one of two reasons: 1) the cost of creating that sample was considerably lower than the cost of creating either a (simple) random sample or coding the population (i.e., they were convenient, which in many cases is synonymous with feasible); or 2) the situation of interest presented difficulties for the enactment of any procedure that would be either random or comprehensive (e.g., war followed by authoritarianism). Regardless of the context, the point should be made that convenience samples are biased samples.

In producing and making public data, researchers should state clearly whether they have produced a sample and report whether it is a random sample or a convenience sample. In

short, the User Guide for conflict data sets should state prominently whether the sample is representative and how that was assessed. When the data are, or may be used as, a convenience sample, the User Guide should contain some sort of discussion about the authors' beliefs about the process that produced the convenience sample. Doing so is the first step to providing information about the nature and degree of the potential bias. In some cases this may be easy; in other cases, the bias may vary across the observations due to differences in temporal and spatial data collection. For example, we would expect the content of news wire reports to be different across countries (i.e., we get more information from Ireland than Senegal) and across time (i.e., there generally are more news reports available the more recent the period; there are more news reports when certain areas are on the international agenda, etc.).

Lastly, the User Guide should encourage users to model the bias, to the extent that is possible, in their analyses. We have included below a brief section about analyzing data which contains discussion that may help researchers write that portion of their User Guide. Very broadly, however, the User Guide should be explicit about how the nature of likely bias may or may not impact the conclusions users would draw in their analyses.[11] For example, in some cases, the nature of the bias may attenuate, rather than bias, the results in favor of a hypothesis. We recognize that users will often devise uses for the data that had never occurred to the data collectors, and that the data collectors cannot, thereby, list all of the potential biases as those will vary considerably across the use of the data to address various research questions. A discussion, however, of the potential bias with respect to research questions that the data collectors anticipate users might probe with the data would be a considerable improvement over the status quo (circa 2014).

Note that research projects that refer to a population are not encouraged to engage in this in part because these types of datasets do not have the problems being identified.

[11] See, for example, the "Health Warnings" discussion in (Bueno de Mesquita, Fair and Shapiro, 2013).

2.3 The Unit(s) of Observation

Collect data at the most fine grained units of observation that the source permits, and aggregate up.

This is arguably the area where we are weakest as a community. Consider how many articles and books you have read in which detailed discussion exists about the temporal and spatial units of aggregation used? We often use spatial units such as "country," "region," and "district." Why? The discussion tends to be quite limited, especially with respect to what causal mechanisms and processes might operate at one such unit, but not another. Our explicit discussion of temporal units is not much better: whether a project produces data over days, weeks, months, years, etc. is rarely discussed, but almost never discussed with respect to mechanisms and processes. A reader might object that these are theoretical, not data, issues, but we push back: those of us who collect data should do our best to not limit the theoretical reach of the data. That is, we should endeavor to make the data as useful to as broad a set of theories as we can.

To that end we advocate a standard that all data collection efforts conduct all coding (whether automated, human or computer-assisted) at the most fine-grained level of spatial and temporal aggregation that a given source permits. This is actually something that is done already as this is normally what the coder would see. Unfortunately, this is information is later discarded. It is important to note that in some projects a "source" will be something as small as a sentence, while in others it may be as large as a collection of documents. Terminology can be challenging here, so an example might prove useful. Consider a project that is coding war crimes committed by actors, and is using the annual reports of a watchdog NGO to do so. In some sentences in those reports the alleged perpetrator will be identified in very general terms, such as "rebels." Other sentences will identify a specific armed group, such as the Lord's Resistance Army. In such a situation we are advocating that a project code both reports (sentences), using a hierarchical scheme where any specific rebel group

(e.g., Lord's Resistance Army) is nested within Rebel Groups. The key insight supporting this standard is that one can use computers to aggregate to higher levels of both temporal aggregation (e.g., from days to months, from months to years, etc.) and spatial aggregation (from street corner coordinates to a county or a country, from a country to a global region, etc.). The war crimes project example would thus produce fewer cases the more precise the identity of the actor (i.e., there would be fewer cases for which a specific rebel group was named than cases for which "rebel group" is the level of aggregation). This difference is reflected in the source, and we advocate data collection systems that produce data which reflect the variance within the sources that are coded. It will then be incumbent upon users to make decisions about how precise they can be when using the data (recognizing that some sort of bias introduced by using only cases in which a specific perpetrator is named). Any alternative coding decisions imposes upon the user a choice that, depending upon the level of detail provided in the documentation, may not be apparent to the user and can require future researchers to recode the source material.

To illustrate further, imagine a project in which one source contains some reports that indicate the hour and minute as well as the street address of police searches and a second source contains only a quarterly list of protest activity across the entire country, and the researchers plan to conduct analyses using the country-year for their unit of observation. Many researchers would instruct coders to record the country and year for each observation, and add up the number of police searchers and protest events. They might spend tens or even hundreds of thousands of dollars on the coding, and a future researcher who might want to estimate a multi-level model using both the weekly number of police searches by province (level 1) and the quarterly level of protest activity per country (level 2) as independent variables would have to raise a similar level of funding and return to the source material to recode it. However, had the first researcher coded the time of day (e.g., using

the 00:00-23:59 time format) and the Geographic Information System (GIS)[12] coordinates for the street address for the police searches, and then the quarter-year and any standard country code (COW, ISO, etc.), that researcher could have then written code in any statistical software package to aggregate the data to the country-year unit of observation and have produced a valuable resource for the community. The GIS research community has produced a remarkable resource that makes is possible for anyone with a computer and a basic knowledge of GIS to aggregate and/or merge files across the spectrum from a jarringly precise latitude and longitude point through countries, or even regions, of the world.

Those who prefer coarser units of time and many spatial cases (e.g., a country-year data base) tend to be interested in cross-national comparison and mean effects. They are confident that the process they are studying is sufficiently uniform across the countries in their sample that the mean effects represent typical, or modal, effects. They are also confident that processes, however they unfold through time, can be modeled either as if they do so annually (as in a country's military expenditures), or as if the process is in equilibrium at the annual level (e.g., a government's response to dissent is consistent throughout the year, such that the mean response over a given year represents the typical state response). These assumptions are almost always implicit, and we would do well to be explicit about them.[13]

Those who prefer more fine-grained temporal and/or spatial units of observation tend to reject the homogeneity of processes as the country-year level. For example, Phil Schrodt has been known to claim to be suspicious of any study that includes both China and Fiji in its sample.[14] To the extent that researchers believe that a government's response to dissent varies substantially and systematically across a given country and/or year will want to model

[12]See the GIS wikipedia entry here: `http://en.wikipedia.org/wiki/Geographic_information_system`. There is more than one potential system for classifying space (e.g., Global Adminstrative Areas and Global Administrative Unit Layers), but it is relatively easy to translate across these systems.

[13]It is tempting to consider standards for data analysis and reporting that the community would adopt, such as a discussion of robustness checks, but that is outside of the scope of this document.

[14]He claims to be equally suspicious of any study that includes only China and Fiji.

that variation. This cannot be done using a country-year unit of observation.

To date, the literature has been poor at engaging debates about these choices. As our sophistication grows, we need to strive to do better making explicit our theoretical beliefs (assumptions) and engaging one one another in debate about the spatial and temporal units over which we claim that the processes we are studying operate.

2.3.1 Space and Time

Since all data that conflict researchers collect have both a spatial and a temporal domain, we should collect both at as fine a level as the source permits, regardless of whether we plan to use the data at a level where there is spatial and temporal variation. While most of us are comfortable with different temporal units of aggregation, relatively few of us are comfortable with different spatial units of aggregation. As a field we need at a minimum to record alpha variables that indicate a spatial location, but ideally record either a central point or collection of points. When that information is recorded it is a matter of making use of well standardized GIS files to aggregate to higher spatial units. Weidmann, Kuse and Gleditsch (2010) have created a GIS file (CShapes) that contains all of the countries in the world between 1946 and 2010, and Tollefsen, Strand and Buhaug (2012) have proposed a GRID based standard that permits inclusion of subnational data, but is compatible with CShapes and provided software.

Should we focus our collective labor upon time-series case studies (i.e., fine-grained temporal units of observation with one or a few spatial obersvations), collecting many data points on a large number of variables, or (pooled) cross-sections (i.e., coarse tempora units of observation with many spatial observations, collecting many data points across different cases? The obvious answer is: both. Yet, our labor is finite and the opportunities infinite. So how do we collectively apportion our labor to maximize cumulation of knowledge?

To date data collection has been skewed strongly toward the latter: the country-year is

the dominant spatial-temporal unit of observation in conflict studies, and both are coarse units. A pioneer in the field, Pitirim Sorokin (1957) studied the civilization-century as a unit of observation, and Charles Tilly (1984) studied more than 11,000 arrest records, nested within demographic and other data collected at both the neigborhood and city levels. That is, despite the dominance of the pooled, cross-national time-series data structure in the statistical study of conflict, there is a long history in the field of researchers using both more fine-grained, and more coarse, spatial and temporal units of observation. Elsewhere in this document (Units of Observation) we discuss the Standards and Practices to employ when collecting data. Here we emphasize the most obvious trade-off, noting that no focus is inherently preferable to another (though for specific theories, given granularity of spatial and temporal units will be preferable).

Spatial units are increasingly embracing a GIS-compatible coding structure, but politically defined spatial units (District, State, Country, Region) remain dominant. Any politically defined unit can be defined using GIS-compatible coding schemes, and we encourage researchers to move away from country-codes (such as COW, Polity, ISO) in favor of GIS-compatible codes (e.g., Weidmann, Kuse and Gleditsch, 2010; Tollefsen, Strand and Buhaug, 2012). The first step in such a transition is to include both country-codes and GIS-compatible codes.[15] The Conflict Consortium will work to create resources to facilitate this important transition.

That said, spatial aggregation is not as "worry-free" as temporal aggregation is in our field. The well known "boundary problem" (`http://en.wikipedia.org/wiki/Boundary_problem_(in_spatial_analysis)`) and related "modifiable area unit problem" (`http://en.wikipedia.org/wiki/Modifiable_areal_unit_problem`) need to be considered. The first concerns aggregation over fixed units, such as a grid. The second involves the changes over time in the types of units that do not suffer from the boundary problem and will often

[15]Indeed, there is no reason to exclude what will eventually become legacy codes.

be of interest to political scientists, such as municipalities, counties, and other political units: these tend to change over time. Users of data must be made aware of these issues, and we promote as a best practice the release of data at the most finely grained spatial unit available (given the source), which is to say in a point-format. Data collectors will often want to also release spatially aggregated data as well, but their User Guide should discuss these issues.

The temporal units of aggregation used in the field range from minute, hour, day, week, fortnight, month, quarter, year, decade, score, and century; however, the year is by far the most common, and the farther one moves in either direction away from year, the fewer data sets one finds that adopt that unit of aggregation.

Scale in Space and Time

Scale refers to a continuum over which one can move up or down a level of aggregation. Conflict scholars are interested in both spatial and temporal scale (noted above), though there is little if any discussion of the issue. The temporal aggregation list above (from the *second* through the *century*) is the scale over which scholars might choose. Because the appropriate scale depends on the question at hand, we can repeat the primary standard and practice for time: all data should be recorded into an electronic record at the most fine temporal scale that the source permits. The project may want to release data using the month for the unit of temporal aggregation, but have some sources that provide information at the day. The initial electronic coding should be done at the day, and then a computer program should be written that aggregates those data to the month. Doing so will permit later researchers (and, in some cases, the PIs themselves) who want to work with that same information source, but at a daily temporal aggregation, to make use of the coding project rather than have to go back to a new content analysis of the original source.[16]

[16]Some may note projects that rely upon different data sources frequently code information across multiple scale levels. To be clear, we are advocating for doing the coding at the finest scale that each distinct source permits, and then using computers to aggregate those up to a common scale. While doing so may seem cumbersome, it increases reliability (humans are not aggregating in their heads), and while the final data

The spatial scale created by GIS is less well understood by researchers in this area, largely because we do not use it in our daily lives as we do the temporal scale. But it is critical that we come to understand that space scales just as time does; that is what Geographic Information Systems (GIS) was created to do. The ability to scale is very powerful as it permits researchers to merge data from a remarkable array of sources and to move up the aggregation scale when doing so. That is why it is so important when planning a new data collection effort to give consideration to how your data can be made GIS-compatible. Again, the Conflict Consortium plans to assist development resources to facilitate this.

In addition to geographic tools there are also variety of measurement models that are capable of linking diverse pieces of information in theoretically meaningful and transparent ways. These measurement models are sometimes referred to as item-response theory models (IRT), factor analytical models, latent variable models, latent class models, or hierarchal models, among many other names. Though the terminology varies, each model shares a focus on linking data to theoretical informed parameter estimates and are actually all quite similar. A note of caution is in order: just as one can throw the "kitchen sink" into a linear regression model, so can one throw the "kitchen sink" into a measurement model. However, just because it is possible to throw a lot of unrelated data into a measurement model does mean that these models are not useful tools for aggregating and linking disparate sources of conflict data. One might ask why use one of these more complex measurement models in the first place? Why not just use a simple model like an additive scale when combining information at different temporal or spatial scales? It is important to note that there is no model free way of combining different pieces of information (Schnakenberg and Fariss, N.d.). Every approach makes explicit assumptions — stated or not — about the way in which the different pieces of data are generated and how the data are related. For example, an " additive scale approach is

will be released to users at an aggregated level, the project should *also* make available the data collected at the finest scale that was used to produce the final data. For an example, see the ITT website's Under the Hood section.

a *model* assuming equally weighted indicators and no error" (Schnakenberg and Fariss, 2014: 13). We wish to reiterate that every measurement model is based on a set of assumptions which produce an estimate that is conditional on data at hand. It is up to the researcher to specify a theoretically appropriate model. We recommend that researchers think carefully about the assumptions they are making no matter the complexity of their measurement model. This suggestion goes along with our other suggestion of disaggregating data collection as much as possible. When there are questions about alternative parameterizations then model comparison statistics can be used to adjudicate between competing models. That is, although estimates from measurement models that aggregate different pieces of information can be evaluated, pre-aggregated data cannot be evaluated in a systematic way (unless of course another researcher wishes to replicate the data collection effort).

The takeaway points, then, are:

1. It is easy to use computers and measurement models to move up in scale, but it is impossible to move down in scale.

2. Both space and time are scalable, and all conflict takes place in *both* space and time.

3. Data should be recorded at the finest spatial *and* temporal scales the source permits, regardless of the place on the spatial and temporal scales that the researcher plans to use the data.

In principle these changes are easy. To put them into practice will sometimes require more resources and each project will have to weigh any such costs against the gains to cumulation. But when we review proposals for funding agencies we should push hard for this standard, noting when relevant that the PIs may require more funds than they have asked for to meet this standard.

2.4 Sources

Provide sufficient detail about source material to permit replication of your data.

There is a general problem with regard to sources (i.e., the primary documents used within the data collection process). For example, it is not quite clear where they get the raw information from, how people code what they do, and what else could be done with the material allowing for the fact that interests might change.

The key appears to be providing all information that was employed in creating data (i.e., compiling an archive), disaggregating this information as much as possible and providing detailed citations that include the title, volume and page number from sources.

2.4.1 Source and Perspective Bias

State whether the data collection seeks to produce a record of the True State of the World, a record of information available at given location and time, or something else. When it is not the True State of the World, a statement about the beliefs about bias should be included.

There is a constant issue with regard to the "Rashomon effect," or source bias problem, whereby individuals attempt to understand what has happened through the use of multiple sources, acknowledging that the sources might communicate different things (Davenport, 2010). Too little consideration has been given to the following questions. When compiling conflict/violence data is one identifying Truth (i.e., an accurate and comprehensive listing of phenomena, processes, etc.)? If so, is one able to produce a census, or is one having to produce an estimate via sampling? If the data collector is not producing a record of The True State of World (in a given spatial-temporal domain), then is she providing a record of the information that was widely available to actors at the time that they were taking action? What other alternatives might the data collector be coding? The answers to these questions must be theoretically guided: we do not collect data in a vacuum. Yet, the existing practice

is to proceed as if we are collecting a census of The Truth, and we fail to discuss the issue in explicit terms. Some examples will help.

Consider first a data collection effort to produce a dataset that will support the prosecution of alleged crimes against humanity. Such an effort might need to establish the number of victims of a crime, such as summary execution of non-combatants. A census of that number is, for a variety of reasons, infeasible: whatever number is produced will contain uncertainty about the True integer value. As such, the effort involves inference (estimation of an unknowable from known information), and statistical inference is an excellent candidate (other valid inferential options may exist, or be developed in the future).

Second, researchers might be studying the decision making of actors in conflict, such as whether they should use non-violent or violent tactics, whether to continue fighting or negotiate cessation, whether to commit/permit acts of mass-killing of non-combatants, or the impact of political institutions upon any of the foregoing decisions. It should be apparent that the True integer count of any type of event, as well as the True ordinal level of any standard based variable, are not of obvious value to a data collection effort that will support hypothesis testing of theories developed to explore these research questions. What is germane is the set(s) of information available to the actors involved in making the decisions about tactics, whether to fight, whether to commit non-combatant slaughter, etc. In other words, any research project that theorizes about the beliefs of actors will want data that provide a valid and reliable summary of the information that would inform such beliefs. For such an effort a defensible case can be made that a census approach is feasible (we make this case elsewhere).

Third, while we are unable to think of data collection projects where the theory to be evaluated using the data would lead one to want to collect something other than information sets informing beliefs, we are confident that other researchers will do so. To that end, we will create both wiki and blog space at the Consortium website where researchers can

contribute their arguments. What this discussion produces is a recommended standard where we abandon our practice of silently assuming what we are collecting conflict data about in favor of explicit discussion.

Another way to think about this is to recognize that virtually all information that we use to produce data was itself produced by a non-random process, and rarely does that information record "ground truth" (i.e., a census). We have a tendency to (1) fail to explicitly theorize about or describe the process that produced the information we code, and (2) when we code all of the information that was available to us, we tend to present our data *as if* it represents a population (or can be treated as if it were a random sample from some population). The field needs to stop these fictions.

We advocate a standard and practice of explicitly stating whether the data collection seeks to produce a record of the True State of the World (for a given spatial-temporal domain), a record of information available (for a given spatial-temporal domain), or something else (for a given spatial-temporal domain). Further, for data that do not claim to represent a True State of the World, the authors should include a statement the describes their beliefs about bias.

2.5 Reliability

Perform and report formal reliability assessments.

A set of operational rules (i.e., a measure of a concept, or a variable) is reliable to the extent that it produces the same value when used to measure the same case (or observation).[17] Social scientists who study conflict do a better job assessing the reliability of the instruments they create to measure variables in their data sets than they do assessing the validity of those measures, but our community is far from uniform with respect to this practice. The appropriate standard is formal reliability tests (e.g., test—retest, interrater reliability scores,

[17]For a good online source, see the Research Methods Knowledge Base's Reliability entry (Trochim, 2000).

etc.). In lieu of formal assessments current practice often substitutes a discussion of the extent to which coders meet with PIs or emphasizes the experience of a given coder. Some data projects do not provide any discussion of reliability (e.g., Polity II, III and IV; `http://www.systemicpeace.org/polity/polity4.htm`).

A related issue concerns the recruitment and training of coders. A training period prior to coding is essential, and coders should not code until they have passed an assessment demonstrating that they are able to produce sufficiently reliable and valid data. This process should be described in detail when data are released.

Similarly, when automated machine coding is performed, reliability analyses should be conducted and reported, and the source code should be made available. Examples include, but are not limited to, machine coded content analysis (e.g., Gerner et al., 1994), machine coding of image files (such as video files or satellite imagery), and other machine (assisted) processes of measurement.

2.6 Validity

Perform and report formal validity assessments.

The validity of a measurement concerns the extent to which the values it produces across cases reflect the actual values those cases have for the concept being measured.[18] To assess validity, then, we must begin with a precise definition of a concept, and that definition must include the range of values specific cases might take for that concept. Conceptualization thus includes some attention to the degree of variation we might observe over that concept (where degree can be understood as nominal, ranked, etc.). An operational definition (i.e., set of instructions for recording a value for a case) is more or less valid depending upon how strongly it reflects the true values for that concept across a given set of cases.

Two standards and practices warrant note. First, social scientists who study conflict

[18]For a good online source, see the Research Methods Knowledge Base's Sampling entry (Trochim, 2000).

tend to give short shrift to conceptual and operational definitions. Researchers generating data should have explicit, precise conceptual definitions as well as an explicit (and therefore replicable) statement about the operational definitions. All projects should contain explicit statements of each.

Second, the community of conflict scholars invest limited effort in formal assessments of validity. Several opportunities exist (e.g., construct validity, content validity, criterion validity), and all data collection projects should undertake, and make publicly available, at least two formal assessments of the validity of the measures produced in the dataset. That said, a narrow focus on a statisical measure of validity that shortshrifts theoretical assessment is also unwarranted.

2.7 Competition for Funding

Researchers on the topic can have a difficult time obtaining funding for their projects. To the extent that we can, as a community, limit the number of times different research teams return to the same source material to generate data, we can ease that competition. The Unit of Observation standard discussed above will contribute to reducing the need to return to source material, thereby reducing demand on scarce funding resources.

Sometimes referees' / funders' conceptions that relevant information cannot be collected undermines support for a project. When that is the issue, a pilot studiy that demonstrates the proof of concept is very helpul.

2.8 Institutional Review Boards (IRBs)

IRBs are sometimes a problem for those of us who study conflict and violence.[19] We know of cases where reserchers were not allowed to engage in the work due to (mis)perceived dangers.

[19]This is true not only of university level IRBs, but also human subject and other requirements imposed by some funding sources.

Further, there appears to be limited standards across IRBs for evaluating the type of work members of this community conduct.

We urge all researchers working on this topic to adopt for their projects the International Committee of the Red Cross's Professional Standards for Protection Work as standards for the conduct of any project personnel or human subjects.[20] Adoption of this comprehensive set of internationally recognized standards should greatly assist human subjects approval.

In addition to promoting this standard, the Consortium will create both wiki and blog space on its website where researchers can share both successful and unsuccessful IRB applications, as well as advice and experiences.

2.9 Recruiting Labor

There are a series of labor issues that arise in the compilation of political conflict and violence data concerning wages, rights, access and contributions to research.

It is generally believed that the best way to address any labor issues is to identify the range of possible employment experiences from volunteerism to wages with full benefits. For example, the Ill Treatment and Torture project (`http://faculty.ucmerced.edu/cconrad2/Academic/ITT_Data_Collection.html`) paid both the PhD student managing the project and the undergraduates doing the coding wages considerably above what they were able to earn at for similar work. The Organization data of the National Human Rights Institutions Data Project (`http://bit.ly/14thvwx`), on the other hand, was collected using strictly volunteer labor (some students earned independent study credit, but many did it strictly for the experience and to build their resumes). The ITERATE data project is a widely used effort that is self-coded by the PI. That said, whatever approach to labor one adopts to produce data, it is critical to be equally respectful of all labor involved.

[20] See, especially, chapter one, "The overarching principles in protection work," and chapter 6, "Managing sensitive protection information."

2.9.1 Ownership

Data collection is rarely an individual enterprise, and as such PIs should clarify at the moment anyone joins a project what stake that individual has in the product. In other words, from the beginning of the effort ownership of the data (including the ability to analyze it, re-release it, etc.) should be clear to all who are involved in its production. A visit to the Creative Commons website might prove fruitful as researchers may want to invoke a Creative Commons license.

2.10 Proprietary and Sensitive Data

Seek an extendable non-disclosure agreement, or other innovative solution.

Sometimes researchers gain access to sources that they are only able to use to collect data under the proviso that they do not share the data. Such an arrangement is at odds with the replication standard and to that end should be avoided. However, individual researchers will evaluate such situations on a case-by-case basis, and some will decide to accept such an agreement.

We encourage scholars who gain access to data only after signing a non-discolsure agreement (NDA) to write conditions into those NDAs that allow the data to be shared for replication purposes with other scholars who are willing to sign onto the NDA. The US government's Bureau of Economic Aanalysis has a Special Sworn Researcher program ¡`http://www.bea.gov/about/research_program.htm`¿ that researchers should evaluate as a model. The Conflict Consortium might be able to create a board of auditors who could be granted access to the material and then sign an affadvit certifying that the data are as represented in research. These sorts of arrangements are also frequently brokered in the real estate and corporate finance literatures. Rather than reluctantly accept the restraints those with proprietary data wish to place upon us, our community needs to become more aware of

alternatives that protect legitimate concerns, but also allow some access to replication.

To the extent that researchers fail to reach agreements that permit them to share replication data in some way, we encourage scholars to discount such data collection and work accordingly. That said, it is critical to point out that there is nothing inherently weak or poor about data or studies that rely upon proprietary sources. However, our inherent skepticism should be heightened, which is to say we should have less confidence in findings from data based on proprietary sources. Rejecting such findings out of hand, however, is inappropriate and should be challenged as such.

In addition to these general points, it seems worthwhile to mention that questions and disputes in this domain could be brought to either the American Political Science Association's Ethics Committee or the American Association for the Advancement of Science's Committee on Scientific Freedom and Responsibility.

2.11 Are Qualitative Data Collections Distinct?

The standards and best practices contained herein apply to any and all data collection projects about conflict. We recognize, however, that some will maintian that there is a meaningful distinction to be drawn. We encourage folks with that view to consult Moravcsik (2014) and visit the Qualitative Data Repository (`https://qdr.syr.edu/`) project.

2.12 Clear Documentation

Produce a written coding protocol. Consider the end user of your data.

First, create *both* a document that is for your coders and a distinct document for the users of the data (i.e., a User Guide). Verbal (or online) training is imperative, but it is critical that the rules for assigning values to cases for each variable be recorded.

Second, from the very beginning of your project make plans for the User Guide. A

standard problem we all face when writing is recognizing what information we have in our brains that we have failed to share with our readers. This problem is magnified when the mind-dulling repetition involved in data collection has burned into our brain the specifics of how we went about it. There is precious little benefit for excellent, and limited cost to poor, documentation of the process. When writing the User Guide it may prove helpful to imagine first year Ph.D. students as the audience. Both the User Guide and the Coding Rules used by coders should be made publicly available.[21]

Third, keep a research journal of all the questions coders ask. This is especially important for multi-year projects as it will help you build a FAQ for future coders (and may well prove useful for drafting and revising the User Guide).

Fourth, bear in mind that the standards are transparency and replicability. The discussion in Sundberg and Harbom (2011, pp. 107-8) is useful.

Fifth, we encourage all data collection projects to "test drive" their User Guide by handing it to someone who was not involved in the project and asking that person to code a dozen or so cases, and having done so, provide the project with feedback about what needs to be improved.

[21] For an example, see the ITT website's Under the Hood section.

3 After Collection is Completed

3.1 Public Release of Data

Release and archive your data on the Internet at a site you do not need to maintain.

The replication standard requires the public release of data. Some researchers do not do so, and this is a problem we must strive to eliminate.[22] Yet, when should researchers make their data available to the public?

Researchers who collect data invest an enormous amount of time in planning, securing funding, recruiting and training coders, cleaning and documenting the data, and then hosting it on a website. This work produces little professional benefit. As such, most researchers who collect data do not release it until after they have had an opportunity to produce their own research with it.

We urge researchers to release their data as soon as it is cleaned and documented. We believe that researchers readily understand the advantages of an embargo period, but do not necessarily appreciate the advantages of immediate public release. For example, Gleditsch, Metelits and Strand (2003) show that articles that include links to the data used are cited

[22] We recognize that some data collection efforts face issues that most do not. See the sections on Proprietary Data / Sources and the one on IRBs.

more often than articles which did not make the data publicly available.[23] Further, the pace of research is changing. Data will increasingly become useful to the extent that they can be merged with other data sources. Making one's data available frequently leads to its use in studies one had never anticipated, and while the individual professional benefit to the researcher tends to be diffuse, it can be quite high. More importantly, the contribution to science cannot be anticipated in advance.

For those who choose to temporarily embargo thier data, we do not believe a one-size fits all standard and practice is best. Individual researchers must first comply with whatever rules govern their project given funding sources. Within those parameters, researchers may want to wait one, two, or perhaps even five calendar years before making the data public. That might be from the data of acceptance of the first publication using the data, or perhaps from its publication date. Anything beyond five years seems unlikely to be defensible.

3.1.1 Archiving

To make data publicly available we must archive it. It goes without saying that all researchers should have a professional website that contains links to their data projects. A question arises, however, where that data ought to be archived. Though none of us likes to think

[23] For those unfamiliar with this discussion, we quote their motivation in full:

> Several scholars argued that a replication policy would prevent scientists (particularly younger ones) from making full use of their own data before making them available to others. Herrnson (1995) claimed that the example of researchers who collected data and were subsequently denied its maximum use would deter scholars from investing the resources or shouldering the risks associated with data collection. Forcing individuals to archive their data sets, he argued, would not increase readership. Original research questions would increase interest in the field and this would require researchers to collect original data.
>
> In stark contrast to this view, proponents of replication have argued that those who release their data will find that others begin to take an interest in their work, and to cite them, and that this will be to their own advantage (King, 1995:446). Although the call for replication was mainly formulated for the good of the overall research community, those favorable to replication tend to argue that the alleged conflict of interest between the interests of the individual researcher and the interests of the scientific community is a straw man. This article is an attempt to test these two contrasting views (p. 89).

about our own demise, our datasets will likely last longer than we will, and for that reason alone, our personal website is not the optimal location for hosting our data.

The Dataverse Network (`http://thedata.org/`), Grow^{up} platform (`http://growup.ethz.ch/`), Interuniversity Consortium for Political and Social Research (ICPSR, `http://www.icpsr.umich.edu/`), and the Qualitative Data Repository (`https://qdr.syr.edu/`) projects are four options where one can outsource the archival problem. The "problem" is not only our limited time on this cold orb, but also that electronic storage media change (consider: punchcards, magnetic tapes, hard drives, floppy disks, removable hard disks [Zip Drives!], USB flash drives, the cloud). The problem arises because once one makes a dataset public, one must maintain the computer system and media on which the data are stored. Researchers who produce conflict data should turn that effort over to an organization that has professional staff and resources dedicated to that effort.[24]

A secondary benefit of depositing data in a publicly accessible archive is that it becomes a focal point for scholars searching for data. The Conflict Consortium will build an eventually comprehensive catalog of all conflict databases, and we also plan to develop an archiving capacity.

With respect to what should be archived, all source material (barring copyright issues—see below) should be scanned to digital form and archived. Note that URLs and other digitial links are not permanent. Hence, the best practice of creating a digital record of source materials rather than only the link.

A growing community is turning to what is known as VCS (version control system) and using platforms like gitub (`https://github.com/`) or Open Science Framework (`https//osf.io/`).[25] The idea of these systems is that all the code and data are composed and

[24]We should note that it is possible to restrict access at many publicly available archives. For example, the dataverse site (`http://thedata.org/`) permits one to password protect files that the PI does not want to be publicily available, but nonetheless wants to archive.

[25]Github in a nutshell: `http://git-scm.com/book/en/Getting-Started-Git-Basics`.

edited on the system (like dropbox but with many more features). The system keeps track of all the changes as they are made in real time so every decision made by a coder, PI, or any other participant in the project is documented. This type of system really takes replication to a whole new level. A study done from start to finish on Github is not only reproducible—in the plug and play sense that is the minimal level of replication now required by most journals—but replicable in that all decisions made to data, code, and writing are documented instantly. Using a site like Github also greatly reduces the cost of pulling all of the relevant replication files and data sources together once a paper is finally "in print" or forthcoming. The only task left once a project is done is to make all the files public if they weren't public in the first place.

Lastly, we acknowledge that social media data presents some unique difficulties here as it is not often clear exactly who owns what and what could be done with diverse material. The best practice here would be to identify what has been done with data that comes closest to the one that is being compiled, finding the practices of those using similar data, contacting the researchers involved and asking them what they did. Additionally, all of the decisions made along the way should be posted for individuals to follow - again back to the community.

3.1.2 Proprietary Sources

As ownership of digital content increases purchasing data may become increasingly important. We already subscribe to services such as Factiva, LexisNexis, etc. via our institutions, and those contracts largely prohibit the public release of source material harvested from them. We anticipate that this is likely to continue and encourage researchers to form consortia that are able to create agreements with proprietary source providers for the sole purpose of replication.[26]

For example, the Conflict Consortium, perhaps in cooperation with a coalition of univer-

[26]For an assessment of the situation circa 2014 see Phil Schrodt's "The Legal Status of Event Data" `http://asecondmouse.wordpress.com/2014/02/14/the-legal-status-of-event-data/`.

sity libraries, might negotiate with Factiva to produce a password protected section of the Factiva site where researchers who code data from downloaded stories deposit those stories. Only consortium members would be able to access the site, and all they would get are the source materials used by the project.

That said, we hasten to add that there are two issues to consider when one's data have been collected from information contained in proprietary sources. The first is the replication of the data itself. Legal ownership (copyright) may prevent this, though as noted in the previous section, we urge authors to pursue extendable NDAs. Yet this is a second issue: transparency. We encourage authors who work with proprietary data and are unable to secure an extendable NDA to make a concerted effort to make the data collection transparent. For instance, clear documentation (discussed below), illustrated with small samples of actual (perhaps redacted) information, can go a long way toward meeting transparency when replication cannot be met.

4 Updating your Data

It is common for data collection projects to release updates. This raises some important issues that have received limited attention. First, a legacy copy of each release of a dataset should be archived and made available to researchers. Why? Replication of studies that used a legacy copy require it.

Second, data producers should make publicly available a record of the revisions made to cases. This might be done in a variety of ways.

- Via a dummy variable coded 0 for observations that have not been revised and 1 for those that were.
- Posting computer code that makes the changes (e.g., a Stata or R command file).

- Posting a written description of each observation for which a variable received a new value.

Ideally the project will make available an explanation for why a value was changed (this might be a note in a cell of a spreadsheet, described in a written document, etc.).

4.1 Version Control and Change Logs

We offer two points of elaboration to best practice with respect to archiving legacy copies.

- Every release of a data set should have a unique version number
- That version number should be a variable which is on every observation in the data set (so a column in every data set which is the "version" variable

Note that VCS (such as github, discussed above) make this considerably easier by making it part of the workflow.

4.2 Cost Concern

Not all projects receive funding: many efforts are self-funded, or receive only partial funding support. The Political Terror Scale, Banks' Cross-National Time-series Archive, and IT-ERATE are prominent examples. Because governments generally have a greater interest in preventing the type of data collection our community conducts than funding it, resources are scarce. As such, there is some concern that creating more demanding standards and practices may lead some projects to cease, due to an inability to meet the standards and practices. This is an issue we encourage people to discuss on the Data blog space.

5 Analyzing Data

Though this document is about producing data, the emphasis on distinguishing populations and samples has important implications for how we analyze data, and therefore we have a brief discussion of the challenges the community faces when analyzing convenience samples. In a document like this one can only paint with broad strokes, leaving detail about specific methods, modeling strategies, etc. to fora dedicated to those tasks.

5.1 Modeling the Data Generation Process

Conflict researchers sometimes do, and often will not, have non-biased samples that represent the population to which they wish to draw inferences. Whether a given data set is representative of a relevant population depends upon the intersection of both the research question and that data being used. To date this complexity has largely been ignored in the field. At a minimum users should speculate (or better, make an empirically grounded / theoretically driven account of) about the data generating process and the likely bias it induces with respect to the inferences they wish to draw in their study using the data.

To return to a discussion from above, a major advantage of conceiving data produced by content analysis of a population of texts *as a population* is it almost forces the user to think about the incentives that would lead the organization that produced the textual record. Doing so is the first step to (theoretically and empirically) modeling why a police agency, news organization, government bureaucracy, or watch dog group would publish some of the information it collects and exclude other, as well as why it might not see some information in the first place. In other words, users will not only have to build theoretical and empirical models of the phenomenon they are trying to study, but also a model of the data generation process of the organization that produced the written material that was coded.

Some population data can be used without concern for modeling the data generation

process, but many efforts will require modeling of the organization's data generation process. Many of the theories that conflict scholars work with (implicitly) contain "information," such as actors' beliefs, as an input to decision making. It can be useful to consider what likely information set a group of people hold in common. News sources provide an arguably valuable record not of "what happened," but instead the information available in the public sphere that shaped widely held public beliefs.

Users might do so by a process known as multiple systems estimation (MSE), a technique that can be used to produce population estimates from three or more convenience samples (e.g., Lum et al., 2010; Gohdes and Price, 2012; Manrique-Vallier, Price and Gohdes, 2013). MSE is complex, relative to current standards and practices, as are many of many other data collection and estimation techniques that have become available as part of the information technology revolution. At a minimum, as a community we need to acknowledge the extent to which the use of convenience samples impacts our descriptive and causal inferences, and ideally embrace and create, methods that permit us to draw valid inferences from convenience samples.

Another approach is to directly model the data generating process, perhaps employing mixture models or other techniques to do so. In 2015, we are in the infancy of such types of modeling efforts, but there is considerable reason to believe that our ability to do this will improve considerably in the medium run.

5.2 Presenting Results

This is a checklist, and perhaps makes more sense as a stand alone set of standards.

- Explanation of your research design: be explicit about how you establish the counterfactual. Answer the question: What would have happened had the cause not been present?

- Show descriptives
- Provide clear explanation of why the partial correlations you show (whether by case study or regression) are informative for the question.
- Provide explicit justification of estimation approach.
- Placebo tests show the reader that things which should have no relationship if your empirical approach is doing its job do, indeed, have no relationship.
- Tell the reader what part of the data the result is coming from, and where it is strongest. If you don't know that, you don't really understand the result.
- Lots of robustness checks.
- Report various functional forms.
- Justify your standard errors

Readers may also be interested in the reporting standards recently endorsed for experimental research (Gerber et al., 2014).

6 For Referees of Funding Proposals

Are you reviewing a proposal for funding to collect data? Please ask whether the PIs explain how they will attend to the following, and to the extent that they do not, consider whether it should be required before the project is funded.

- Theory: clarity
- Literature: has the author demonstrated awareness of related projects, going back in time, and differentiated it from them?
- Conceptualization and operationalization: clarity, disaggregation, replicability

7 Who are we?

The Conflict Consortium (CC) is an NSF-funded collaborative effort led by Christian Davenport (University of Michigan) and Will Moore (formerly Florida State, now Arizona State) but involving scholars from around the world. The effort is put forward to collect and provide information as well as facilitate interaction in a diverse array of venues - both actual and virtual.

Davenport was initially exposed to event data at the nation level through his use of the World Handbook of Political and Social Indicators (Taylor and Jodice, 1983) as well as coding data for the late Arthur Banks Cross Polity Survey. Later, following the suggestions of Charles Tilly, Ted Gurr, Doug McAdam and Sarah Soule, he began collecting his own data, beginning with NSF funded research on state-dissident interactions in the US (Davenport, 2010) and Rwanda (Davenport and Stam, 2003, 2009) as well as more recently India (Macwan et al., 2009), Northern Ireland, Zimbabwe and Mexico. Diverse global databases on security force institutions and large-scale mass human rights violations by the dyad are underway.

Moore was first exposed to data collection as a research assistant to Gurr's Polity II project (Gurr, Jaggers and Moore, 1989). He was also exposed to events data, courtesy of Mike Ward, and read with great interest the events data that researchers from Pitirim Sorokin (1957) through Charles Tilly (Horn and Tilly, 1988; Tilly and Zambrano, 1989) collected. For his dissertation, Moore created an events data scheme modeled on the COPDAB interstate events project and hired a fellow PhD student to conduct content analysis on news reports about the nationalist struggle for Zimbabwe. He then used his dissertation data as a pilot for the Violent Intranational Conflict Data Project (VICDP). That led to the NSF-funded Intranational Political Interactions (IPI) database collected by Moore and David Davis. He has also served multiple terms on the oversight boards of the Minorities at Risk project and the Polity project, and most recently is co-PI (with Courtenay R. Conrad) on the NSF-funded

Ill Treatment and Torture (ITT) project.

References

Baum, Matthew and Yuri Zhukov. 2015. "Filtering revolution: Reporting bias in international newspaper coverage of the Libyan civil war." *Journal of Peace Research* forthcoming.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2015. "Integrated Crisis Early Warning System (ICEWS) Dataverse.". **URL:** *http://thedata.harvard.edu/dvn/dv/icews*

Bueno de Mesquita, Ethan, C. Christine Fair and Jacob N. Shapiro. 2013. *Codebook for BFRS Dataset of Political Violence in Pakistan*. Princeton: Empirical Studies of Conflict. **URL:** *https://esoc.princeton.edu/files/bfrs-political-violence-pakistan-dataset*

Cingranelli, David L. and David L. Richards. 2010. "The Cingranelli-Richards (CIRI) Human Rights Dataset.". **URL:** *http://www.humanrightsdata.org*

Cochran, William G. 1953. *Sampling Techniques*. New York: John Wiley & Sons.

Conrad, Courtenay R, Jillienne Haglund and Will H Moore. 2013. "Disaggregating Torture Allegations: Introducing the Ill-Treatment and Torture (ITT) Country-Year Data." *International Studies Perspectives* 14(2):199–220.

Conrad, Courtenay R, Jillienne Haglund and Will H Moore. 2014. "Torture allegations as events data Introducing the Ill-Treatment and Torture (ITT) specific allegation data." *Journal of Peace Research* 51(3):429–438.

Davenport, Christian. 2010. *Media Bias, Perspective, and State Repression: The Black Panther Party*. New York: Cambridge University Press.

Davenport, Christian and Allan C Stam. 2003. *Mass Killing and the Oases of Humanity:*

Understanding Rwandan Genocide and Resistance. National Science Foundation, Grant SES-0321518.

Davenport, Christian and Allan C Stam. 2009. "What really happened in Rwanda?" *Pactific Standard* October 6:http://www.psmag.com/navigation/politics–and–law/what–really–happened–in–rwanda–3432/.

Gerber, Alan, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R Biggers and David J Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1(1):81–98.

Gerner, Deborah J, Philip A Schrodt, Ronald A Francisco and Judith L Weddle. 1994. "Machine coding of event data using regional and international sources." *International Studies Quarterly* 38:91–119.

Gleditsch, Nils Petter, Claire Metelits and Havard Strand. 2003. "Posting your data: Will you be scooped or will you be famous?" *International Studies Perspectives* 4(1):89–97.

Gohdes, Anita and Megan Price. 2012. "First Things First Assessing Data Quality before Model Quality." *Journal of Conflict Resolution* .

Gurr, Ted Robert, Keith Jaggers and Will H. Moore. 1989. "Polity II Codebook.". Center for Comparative Politics, Department of Political Science, University of Colorado, Boulder.

Hathaway, Oona A. 2002. "Do Human Rights Treaties Make a Difference?" *Yale Law Journal* 111:1935–2042.

Horn, Nancy and Charles Tilly. 1988. *Contentious Gatherings in Britain, 1758-1834*. Inter-university Consortium for Political and Social Research: ICPSR Study No. 8872.

Landman, Todd and Edzia Carvalho. 2009. *Measuring human rights*. London: Routledge.

Lum, Kristian, Megan Price, Tamy Guberek and Patrick Ball. 2010. "Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998-2007." *Statistics, Politics, and Policy* 1(1).

Macwan, Martin, Christian Davenport, David Armstrong, Allan Stam, Monika Kalra Varma and Amanda Klassing. 2009. Understanding Untouchability: A Comprehensive Study of Practices and Conditions in 1589 Villages. Technical report Navsarjan Trust Ahmedabad, India: .

Manrique-Vallier, Daniel, Megan E Price and Anita Gohdes. 2013. Multiple Systems Estimation Techniques for Estimating Casualties in Armed Conflicts. In *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*, ed. Taylor B. Seybolt, Jay D. Aronson and Baruch Fischhoff. New York: Oxford University Press pp. 165–182.

Moravcsik, Andrew. 2014. "Transparency: The Revolution in Qualitative Research." *PS: Political Science & Politics* 47(1):48–53.

Nardulli, Peter F, Scott L Althaus and Matthew Hayes. 2015. "A Progressive Supervised Learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* forthcoming.

Schnakenberg, Keith E. and Christopher J. Fariss. N.d. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods*. Forthcoming.

Sorokin, Pitirim Aleksandrovič. 1957. *Social and Cultural Dynamics: A Study of Change in Major System of Art, Truth, Ethics, Law, and Social Relationships*. New York: American Book Company.

Sundberg, Ralph and Lotta Harbom. 2011. Systematic Data Collection: Experience from the

Uppsala Conflict Data Program. In *Understanding Peace Research: Methods and Challenges*, ed. Kristin Höglund and Magnus Öberg. London: Routledge pp. 91–113.

Taylor, Charles and David Jodice. 1983. *World Handbook of Political and Social Indicators III*. New Haven: Yale University Press.

Tilly, Charles and Lynn Hollen Lees. 1984. *Analysis of Arrests in Paris, June 1848*. Inter-university Consortium for Political and Social Research: ICPSR Study No. 49.

Tilly, Charles and Raul Zambrano. 1989. *Violent Events in France, 1830-1860 and 1930-1960*. Inter-university Consortium for Political and Social Research: ICPSR Study No. 9080.

Tollefsen, Andreas Forø, Håvard Strand and Halvard Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49(2):363–374.

Trochim, William. 2000. *The Research Methods Knowledge Base*. 2nd ed. Cincinnati: Atomic Dog Publishing.

Weidmann, Nils B, Doreen Kuse and Kristian Skrede Gleditsch. 2010. "The Geography of the International System: The CShapes Dataset." *International Interactions* 36(1):86–106.