THE COMMANDER'S DILEMMA
ONLINE APPENDICES

CONTENTS

1. What before why: On descriptive inference with quantitative conflict data

This appendix draws on a 2016 working paper, "What before why: The importance of descriptive inference in quantitative conflict studies."

In summer 2007, I traveled to Colombia with a graduate school colleague to collect data on data collection. During three weeks of interviews, we spoke with several dozen technical staff members at NGO's and government agencies, each of whom was heavily invested in the project of carefully and accurately counting homicide deaths in Colombia, and many of whom believed strongly that their data were more comprehensive, more accurate, and more representative than those available elsewhere. Interview respondents listed dozens of reasons for discrepancies between national homicide lists. Some defined "homicide" differently than others. Those whose work was based on media reports noted their differing monitoring techniques. However, as we noted in our report on the interviews (Hoover and Lynch 2008):

> The answers we obtained were perhaps most insightful in that which they consistently failed to mention: selection bias. Not a single interviewee, when asked why different organizations arrive at different numbers, volunteered that some organizations might be more likely to obtain certain forms of data than others...[I]nterviewees only mentioned sources of difference over which they, or other organizations, had control. They had thought carefully about how they believed the object of measurement should be defined, what methodology they should use, and how they believed political agendas might affect results. But they had not considered forces outside their control that might affect the quality of their data.

These conversations became touchstones for me, for several reasons. First, data collectors were adamant that even major discrepancies between datasets could be explained with reference to easily identifiable (and, generally, controllable) factors, such as definitions, technology, or

political ideology. Second, data collectors often believed that their data were comprehensive, claiming that others missed, duplicated, and/or fabricated events. Third, and somewhat contradictorily, data collectors were aware of significant holes—occasionally including frauds—in most of the data sources we considered. For example, an interviewee employed by one of the largest state data collection agencies informed us that "some" local offices were not counting homicides accurately because of financial incentives for lower (reported) homicide rates. An NGO staff member reported that her organization's office in a strongly conflict-affected area had closed, effectively ending direct homicide reports from that region. Fourth, most data collection organizations in Colombia—like most data collectors in the human rights world more broadly—believed that their data could be easily interpreted.

These attitudes are, I imagine, familiar to most quantitative researchers of human rights and armed conflict. We insist that our dataset is the accurate one, or at least that it "can't be that bad." We identify common types of bias (e.g., "urban bias", "political bias," cf. Kalyvas 2006), but reassure ourselves that the potential magnitude and direction of bias is knowable ("signing" the bias). We default to the assumption that our data are "good enough," employing quantitative data in complex models designed to perform *causal* inference without first investing in *descriptive* inferences. We have no systematic way of determining how inaccurate data can be before our regression coefficients lose significance and/or reverse signs. We don't know what we don't know. In order to make confident conclusions about why or how, we must first understand what.

While there are many types of numerical data in use, I focus here on the type of data that underlie the quantitative analyses in *The Commander's Dilemma*: non-sample event count data

(e.g., killings per department-year). These are often employed as the dependent variable in Poisson, negative binomial (NB), or zero-inflated Poisson or NB regression analyses. In the book, I argue that most regression results based on event-count data are not robust to the levels of bias and missingness that are common in violence data—and that therefore, MSE or other methods provide better estimates of patterns of violence. After discussing what is known about "usual" levels of missingness in conflict count data, I introduce a simple simulation exercise. Working from the replication data of Wood et al. (2012), I simulate a variety of plausible "true" dependent variable distributions, and consider how frequently the main result from Wood et al. (2012) is reproduced using these simulated data. I am not arguing that the Wood et al. data are "bad data." Rather, as with the data from El Salvador that are the backbone of this book, I argue that these datasets, though collected carefully and in good faith, nevertheless require careful, systematic reasoning about *descriptive* inferences in order to form the basis of any *causal* inferences.

First, some terms. In this Appendix, by "convenience data" I mean any data that are not gathered via a systematic sample. Convenience data include media reports, hospital records, NGO records, government reports, and all other types of data derived from passive surveillance. By "count data" or "event count data" I mean datasets that count some type of event across a set of spatiotemporal (or other) strata. Event-count datasets include the number of fatalities (or episodes of protest activity, episodes of sexual violence, etc.) reported via a convenience data source for each country-year (or department-week, municipality-month, etc.) in a given range. The use of convenience count data is accepted practice that leads to publication in high-level political science journals. For example, in a much-read recent special issue of the *Journal of*

Peace Research ("Special Issue on Communication, Technology and Political Conflict," ed. Nils Weidmann, 2015), six of seven empirical articles employ raw event-count data in order to test arguments about the relationship between information technology and (reported) political violence.

However, other studies have demonstrated conclusively that, for many or most conflicts, ostensibly comprehensive lists of violent episodes report only a minority of the events that take place. A key exception to this rule is Bosnia, where excellent pre-war record-keeping and extensive international involvement in a relatively small geographic space meant that nearly all homicides were documented by one or more sources during, or soon after, the conflict (Ball, Tabeau, and Verwimp 2007). Many quantitative conflict scholars assume, implicitly or explicitly, that all cases are "like Bosnia"—but there is little reason to believe this is so. Instead, it is common to see reporting rates as low as 33% (Perú; see Ball et al. 2003), or 40% (Kosovo; Ball et al. 2002, Hoover Green 2010).

Of course, undercounts by themselves do not lead to mistaken inferences. Mistaken inferences occur when incomplete convenience data differ systematically from the true population of events on theoretically relevant dimensions. Systematic differences between reported and unreported episodes of violence can occur for many reasons. In most cases, purposeful manipulation of data plays a very minor role in the differences between observation and reality. The more significant, and complex, source of variation in reporting is the social production of data (cf. Davenport and Ball 2002, Tate 2007). Social production includes both access issues and issues related to networks of trust, knowledge, and reputation; it also encompasses variation in visibility between different categories of victims. A few of the potential

sources of variation in reporting rates (i.e., the proportion of total violent events reported to a given source) across spatiotemporal units include:

- More intense violence may be associated with higher *or* lower reporting rates, depending on accessibility of, and targeting patterns within, affected areas. Reporters may flock to scenes of extreme violence, but reporting by individual victims may be effectively deterred.
- Reporting rates may lag violence, showing no violence when extreme violence occurred and considerable violence in the aftermath of the "main event" when observers arrive.
- Rural areas often have lower reporting rates than urban areas, but variation in reporting rates across different rural and urban areas is significantly greater than variation between the two types of areas.
- Fatalities from large events are much more likely to be reported in media, but affect different populations than single-victim events such as disappearances.
- Changes in logistical situations, such as local support staff and weather, strongly affect reporting agencies' capacity.
- Public figures and high-status individuals are more likely to be reported missing or killed.
- Certain types of violence are more likely to be reported than others, and the likelihood of reporting may be correlated across types of violence. For example, victims may be less likely to report a sexual assault that occurred in private, or a sexual assault that occurred in the absence of accompanying lethal violence.
- Some victims are more likely to report violence than others, particularly taboo forms of violence. Reporting on sexual violence, for example, varies widely across countries and cultures (Palermo et al., 2014).
- Victims and witnesses of violence are more likely to report to agencies or organizations that they trust, implying quite different reporting populations to (for example) government and non-government sources.

In combination, these types of dynamics suggest, frustratingly, that under-reporting of violent events is neither random (as is assumed by traditional regression models) nor simple. If it were random, the potential inferential errors from undercounting would be more knowable (and

more constant across datasets). If it were simple, and/or similar across multiple cases scholars could easily anticipate the direction of bias in a given dataset, allowing them to argue convincingly that (for example) undercounting in a particular count variable produces a conservative bias in the final model. Unfortunately, the assumptions underlying these arguments are generally not warranted.

For example, consider the relationship between convenience count data and rigorously estimated count data for homicides across department-years in Colombia, 2003-2010 (Krüger and Hoover Green 2015). The two largest datasets on criminal homicide in Colombia, each ostensibly complete, do not show complete overlap. In many department-years with hundreds or thousands of estimated fatalities, one or the other of these datasets shows zero recorded homicides. Second, reporting rates are extraordinarily variable across department-years, ranging from about 0.2 to nearly 1. Finally, reporting rates are significantly more variable among department-years with lower numbers of reported homicides. For department-years with 1000 or more reported homicides, the highest ratio of estimated to reported homicides is approximately 3.5. By contrast, analogous ratios for department-years with reported homicide counts of 1000 or less may be orders of magnitude larger. Yet the overall correlation between observed and estimated counts is quite high, above 0.8. That is: a significant proportion of violence is not reported by any convenience data source. It is reasonable to expect reporting rates of no more than 20-60%. However, reporting rates do not vary randomly. In the Colombian case, reporting varies considerably by region.

It is possible to show, using a variety of techniques, that raw data from two ostensibly comprehensive datasets produce notably different results from one another, and from rigorous

estimates. To take one example, replacing one ostensibly comprehensive source of violence data (source A) with another (B), or with estimates derived via MSE, produces quite different results in the same negative binomial regression model of homicide in Colombian department-years. In particular, in a 2016 working paper I found that models with data A reached strikingly different conclusions from those with data B about which variables (if any) policy-makers ought to consider as they determine which areas face particularly high risk of violence, and/or what policy options might limit homicides. Without some benchmark, it would be impossible to adjudicate between these models. Is population the only significant correlate of homicides, as in A, or are homicides associated with political conflict (as indicated by land seizures) as in B? Without accurate descriptive inference, these causal questions are all but unanswerable.

Using MSE data, I found that several factors were significantly associated with homicide rates, including coca production and education, which had not appeared significant in previous models. Colombian homicide data represent, in many ways, an "easy case:" overall homicides are considerably less rare, less politicized, and less stigmatized than conflict-related homicides, and Colombia's data-collection systems are relatively advanced compared to those in many conflicted countries (see discussion in Roth, Guberek and Hoover Green 2011). What are researchers to do when no benchmark is forthcoming? Until recently most researchers have simply treated convenience count datasets as complete, skipping questions about descriptive inference altogether. Here, I outline a basic approach to simulation, then consider results from a simulation exercise based on the model presented in Wood et al. (2012). I note that I have performed similar simulations with other articles using event-count dependent variables, with

similar results. This exercise is not intended as a takedown of a specific article, but instead as an illustration of what assumptions are required to work with convenience violence data, in general.

At the most general level, the question in this exercise is: how often are count regression models robust to plausible levels and patterns of missing data? There are several ways to define and operationalize "plausible levels and patterns" in this context. For the purposes of this exercise, I assume that reporting rates between 0.25 and 0.99 (i.e., rates of missingness between 0.01 and 0.75) are "plausible," and select at random a reporting rate for each iteration of the simulation from a uniform distribution on (0.25,1). For each iteration, I then generate a simulated dependent variable vector by adding simulated "previously missing" observations to the existing dependent variable. How these simulated observations are distributed among spatiotemporal units is a more difficult question. The simplest choice, which is implemented for this exercise, is a fully random distribution of new observations, using random draws from a distribution that roughly matches the overall distribution of the existing dependent variable (in the case of Wood et al. 2012, many zeroes and smaller values, with a long rightward tail), such as a half-normal distribution, a Poisson distribution, a negative binomial distribution, or a zero-inflated Poisson or negative binomial distribution.

More specifically, I used the following procedure:

1. The distribution of simulated data (among spatiotemporal units) is chosen at random from among half-normal (rounding to integer values), Poisson, and negative binomial distributions.

2. From a uniform distribution of plausible reporting rates (in this exercise, 0.25 to 0.99, inclusive), a total reporting rate, RR, is drawn at random.

3. The number of simulated episodes of violence to distribute, NDIST, is derived from the randomly selected reporting rate, and divided by the number of spatiotemporal strata in the existing dependent variable, N, to produce a mean value MU for the simulated observations.

```
NDIST = (sum(DV)/RR)-sum(DV)
NDIST = (1000/0.50)-1000
NDIST = 2000-1000
NDIST = 1000
MU = NDIST/N
```

4. If the distribution type is negative binomial, from a distribution of negative binomial shape parameters (in this case a half-normal with mean 0 and standard deviation 0.75) a negative binomial shape parameter THETA is selected.

5. Simulated episodes of violence, SIMDV, are drawn from a random distribution with N observations, mean MU and (in the case of negative binomial iterations) shape parameter THETA.[1]

6. The simulated variable is created by adding DV to SIMDV.

I repeated this procedure 2400 times, creating 2400 simulated "true" dependent-variable vectors. For each simulated dependent variable, I re-run the model specified in Wood et al. (2012, Table I, model 1), which tests the relationship between one-sided violence by rebel forces, as measured by the UCDP One-Sided Violence Dataset (Eck and Hultman 2007), and a lagged measure of rebel strength, using negative binomial regression with a conflict-dyad-year unit of analysis. I consider a given simulation to have replicated the Wood et al. model when it reaches a substantively similar conclusion (a negative and significant value on the key independent variable *lnintv_ratiolag*, where "significance" is defined liberally as $p<0.10$). Of the 2284

[1] In the case of iterations drawn from a half-normal distribution, the half-normal is a special case of a folded normal distribution with mean zero, folded at zero. Thus the distribution is defined by its standard deviation: `SD = (MU*sqrt(pi))/sqrt(2)`. In R, the distribution is given as `NEWDV <- round(abs(rnorm(N, 0, SD)))`

iterations for which negative binomial regressions converged, just 65 replicated the Wood et al. findings. Of these, the vast majority came from iterations with reporting rates of 0.90 or greater.

In this particular example, then, extremely (implausibly?) high reporting rates, and high correlations between the observed dependent variable and the simulated dependent variable are both necessary—but not sufficient—to replicate the original finding. This is not the case for every convenience count dependent variable. Other findings may prove more robust to lower reporting rates. And we might imagine simulation exercises in which the distribution of simulated episodes of violence over spatiotemporal units has a systematic (model-able!) component. But, as I discussed in chapter 6, it is difficult to determine *ex ante* which factors will influence reporting rates in a given time or place. The appropriately data-critical approach therefore likely involves simulating a wide variety of potential data-generation processes, including randomness and quasi-randomness.

In the absence of simulation (or as a basis on which to build simulations), quantitative researchers must *at least* fully understand the sources of their data, and present their assessment of the potential for variation in reporting rates across units of analysis. This includes considering why some areas may be more or less likely to show up in media reports; assessments about how violence itself may affect reporting patterns; and, ideally, interviews with data collectors or collators. The interviews I helped to conduct in Colombia in 2007 have informed our understanding of variations in reporting across multiple homicide datasets.

2. Methods short of MSE

Unfortunately, in practice, MSE estimates do not converge for all strata we might like to compute for the Salvadoran case, meaning that in some cases data sparseness causes one of three things to happen: either the point estimate cannot be computed, the confidence interval cannot be computed, or the confidence interval includes values that exceed the total population of El Salvador. This is not uncommon; human rights data, especially data from a decade or more ago, are generally so sparse and of such limited quality that even the lower-bound Chao (1987; see Baillargeon and Rivest 2007b) estimator fails to produce a reasonably good-fitting model. For the social scientist whose goal is maximizing both observations and accuracy, the question then is whether a method falling short of MSE can provide useful data. In this section, I discuss the application of what I have termed "overlap analysis" in concert with raw data, as a sub-optimal but potentially useful "MSE fallback." I do not advocate the use of raw data uncomplemented by overlap data, and nor do I advocate making firm conclusions from this method.

"Overlap analysis" has its roots in a white paper by Ball et al (2007) addressing claims about homicide during Colombia's civil conflict. The authors analyzed homicide data, attempting to assess the Colombian government's claim that falling rates of reported homicides represented actually falling rates of homicides. To do so, the authors selected homicide cases at random from a list generated by an NGO, and determined whether the randomly selected cases were also reported in government homicide data. Thus the "overlap rate" is the proportion of all reported cases that is reported on more than one list. Ball and colleagues compared overlap rates between the NGO list and the government list from the periods before and after a key policy initiative. They found that, while government-reported homicides had declined, overlap rates with non-

government homicide lists had also declined significantly. Because an observation of declining overlap is consistent with the hypothesis that reporting had declined while homicides remained relatively constant, the authors determined that the government's claims were premature.

More generally, when overlap rates decline, this observation is consistent with a state of affairs in which the rate of reporting has also declined. Conversely, when overlap rates increase, this may indicate increased reporting rates as well. The table below provides a rough guide to interpretation; in general, the possibilities for inference are strongest when overlap rates and reported violations show opposite trends, and weakest when they match.

Table A1. Interpreting reporting and overlap changes.

	Overlap rate increases	**Overlap rate constant**	**Overlap rate decreases**
Reported violations increase	No inference possible. Could signify increase in violations or increase in reporting.	Weaker evidence of increase in true number of violations.	Consistent with increase in true number of violations
Reported violations constant	Weaker evidence of decrease in true number of violations	Weaker evidence of constant true number of violations	Weaker evidence of increase in true number of violations.
Reported violations decrease	Consistent with decrease in true number of violations	Weaker evidence of decrease in true number of violations.	No inference possible. Could signify decrease in violations or decrease in reporting.

In the Salvadoran case, a significant number of comparisons that cannot be accomplished via MSE can be accomplished, in a more limited way, by assessing raw data and overlap data together. As laid out in the framework above, we cannot reasonably infer from the reported pattern of violence to true variation over time during the earliest years of the war, because patterns of violence match patterns in the overlap rate. However, it is reasonable to infer on the basis of these data that (for example) violence declined between 1983 and 1984, because reported violence declined while overlap rates increased. Similarly, opposing trends in the

number of cases reported and the overlap rates between 1986 and 1987, and between 1987 and 1988, imply that it is reasonable to infer that the patterns in the raw data are correct.

It is important to recall that the relatively constant overlap over time for all reported violence masks significant variation in overlap rates within a number of strata. Much like the comparison between overall overlap rates and variation in the levels of violence over time, it is possible in some cases to make over-time comparisons within individual departments, armed groups, or violation types.

3. MSE Details

Although they carry significant drawbacks, including relatively high costs and biases arising from the relative rarity of wartime violence, surveys are frequently regarded as the conventional "best" choice for estimating casualties during armed conflict. However, in the Salvadoran case, survey investigations of violence were not conducted during the period of the conflict (1980-1992). Nor were surveys conducted in the post-conflict years; at this point, the conflict period is so many years distant that survey research is simply not an option. I have chosen to employ multiple systems estimation (MSE) to produce statistical estimates of violence from four convenience samples.

MSE and related techniques are frequently used to estimate specific population sizes—whether those are populations of animals, of humans with particular conditions, or of casualties. While the execution of MSE for human rights violations can be complex, the intuition is relatively simple: the size of the overlaps between several convenience datasets provides strong evidence about the size of the overall population, including members of the population not counted in any list. In this section, I briefly review the history and theoretical basis of the technique, before discussing model selection, the data-matching process and other practical implementation issues.

MSE was developed in the context of population biology (Petersen 1896), where it has become known as capture-recapture or multiple-recapture analysis (see, e.g., Darroch 1958; Edwards and Eberhardt 1968; Fienberg 1972; Burnham and Overton 1978; Chao 1987, 1989,1992; Pollock et al. 1990; Darroch et al 1993). Its first published application to a human population occurred in 1949, when Sekar and Demings used capture-recapture analysis to correct

census figures for New Delhi. It has been used to determine the population of lesbians in Allegheny County, Pennsylvania (Aaron 2003), the population of drug-abusing HIV victims in Bangkok (Mastro et al. 1994), the population of drug users in London (Hickman 1999), and the incidence of Fetal Alcohol Syndrome in Alaska (Egeland et al. 1995).

MSE was first used to estimate the magnitude of human rights violations in the 1990's, when Ball et al. (2000) used MSE to study killings during the Guatemalan civil war. Since that time, it has been used to estimate casualty counts during several conflicts, for example in Kosovo (Ball et al. 2002, 2007), Bosnia (Brunborg et al. 2003), Perú (Ball et al. 2003), East Timor (Silva and Ball 2006, 2007), and Colombia (Guzmán et al. 2007, Lum et al. 2010). As I discuss below, most MSE applications for human populations (including human rights applications) require at least three separate datasets (or "systems"; hence the name "multiple systems estimation"). However, the intuition behind MSE is best understood by deriving the two-system population estimator and then discussing its four key assumptions.

The basic intuition behind MSE comes directly from probability theory: the probability of being selected in two random samples A and B is simply the product of the probability of being selected in sample A and the probability of being selected in sample B, i.e., $Pr(A) \times Pr(B)$. In a population of size N where all units have equal probability of being captured, the probability of selecting a unit in a random sample of size A is simply A/N; the probability of selection into a sample of size B is B/N. The probability that a unit is selected in both samples (A and B, where the number of units selected in both samples, also referred to as the overlap, is usually called M) is equal to the product of the individual probabilities (A/N and B/N). Thus, if we know how many units are in both A and B (i.e., in M) then we can easily solve for the population size N.

Assumptions However, as readers familiar with probability theory may have noticed, this simple derivation is only true when four assumptions hold. First, the population N must be a "closed system," with no members entering or leaving during the period of measurement. In wildlife studies such as that of Edwards and Eberhardt (1968), this is accomplished by way of a rabbit-proof fence or similar physical enclosure. In the human rights context, it means that casualties do not disappear. While the closed-system assumption can be violated (e.g., when persons thought to be dead are confirmed alive), it does not present a serious problem in practical applications. The second assumption is perfect matching; that is, the relative sizes of M, A, and B must be correctly identified. The practical issues associated with this assumption are difficult but surmountable, as discussed below. However, the final two assumptions of the two-system estimator imply that MSE with only two datasets is usually invalid.

The third assumption is that all elements of the population (all cottontails, all deaths, etc.) have equal probability of capture in any given sample. An influential simulation experiment in population biology (Edwards and Eberhardt 1967) considered a known population of cottontail rabbits (N = 135) released into a four-acre enclosure and then captured by trapping on 18 successive days. In the human rights case, the population might be killings or other acts of violence; individual convenience samples (referred to as data systems or "lists," meaning lists of casualties) serve as analogs to Edwards and Eberhardt's successive trappings. Edwards and Eberhardt's experiment demonstrated the existence of unequal probability of capture, even in seemingly random processes: some rabbits were "trap-fascinated" or "trap avoidant," relative to the general population, across all episodes of trapping; thus, Edwards and Eberhardt found, the true correct number of rabbits could not be reproduced without modeling this unit heterogeneity.

The human rights analog is clear: some victims in some socioeconomic or spatiotemporal or other strata are much more (or less) likely to be reported than others. Unequal probability of capture (the phenomenon characterized above as "uneven under-reporting" or "uneven bias") is a key reason behind the unsuitability of single lists of casualties (or even two lists) as a basis for inference.

Fourth and finally, the two-system model assumes that lists are independent—that being "caught" by sample A does not change one's probability of capture in sample B, or vice versa. However, this assumption is frequently violated. In many cases, non-independence of lists is actually induced by unequal probabilities of capture across (types of) individuals. Darroch et al. (1993) provide the clearest statement of this mechanism of dependence: "Suppose that the lists are independent within strata but the probability of capture or inclusion varies across strata. When the strata are combined, the resulting data will in general no longer exhibit independence" (1138). To be clear, a "stratum" is any defined subset of the full population. Darroch and colleagues, at the United States Census Bureau, typically are concerned with the number of people living in the United States; strata within this population include the elderly, African Americans, immigrants, midwesterners and other key groups across which the probability of returning a census form is likely to vary.

In political violence data, the particular (types of) strata relevant to this dynamic are often unknown ex ante, although it is commonly assumed that (for example) victims in rural areas are less likely to be counted than victims in urban areas. Lists of casualties may also be per se non-independent, meaning that population characteristics are not the factors inducing list dependence. Fienberg et al. (1999) have likened this type of list dependence to "unequal effort" across lists.

Similarly, in wildlife management applications of MSE, system dependences have been modeled as "time effects" (e.g. Chao et al. 1992): the possibility that results may be correlated across proximate capture episodes. Whatever the provenance of list dependence, the degree of dependence, like heterogeneous probabilities of capture, must be parameterized and modeled in some fashion.

Dealing with violations of the assumptions Early implementations of multiple systems estimation generally dealt with (i.e., parameterized and thereby accounted for) only one of the two key assumption violations discussed above, assuming system independence when modeling unequal probability of capture or vice versa (cf. Fienberg 1972). Statisticians have created capture-recapture models that are robust to violations of both assumptions (i.e., that recover correct values in simulations with both unequal probability of capture and list dependence). However, most political violence datasets are too small or sparse to allow for explicit modeling of both these issues. Instead, analysts frequently employ some combination of stratification and parameterization.

Ball and colleagues (2002, 2003, 2007, 2010) typically employ stratification in order to ameliorate issues of individual capture heterogeneity, and focus on parameterizing residual list dependence, a technique formalized in Bishop, Fienberg and Holland (1975). Sekar and Deming (1949) (and many others) have shown that bias due to unequal catchability can be greatly reduced via stratification. Residual list dependences can then be modeled using a log-linear regression in which the "capture" of individual events into individual lists (i,j,k,l) is treated as a Poisson process. With four casualty lists (such as those I employ from El Salvador), data for these regression models comprises the (2^4)-1=15 known cell counts in a table describing

overlaps between the lists, as shown below. The unknown cell count—cases not captured in any list—will be estimated via the intercept term in the resulting model. The "saturated" log-linear model is formalized as follows:

$$\log(m_{ijkl}) = int + u_i + u_j + u_k + u_l + u_{ij} + u_{ik} + u_{il} + u_{jk} + u_{jl} + u_{kl} + u_{ijk} + u_{ijl} + u_{ikl} + u_{jkl}$$

That is, the log of the expected count of cases captured in all four lists (m_{ijkl}) is equal to the sum of the intercept (denoted int above and corresponding to the log of the total number of uncounted cases) the coefficients associated with inclusion in each dataset individually (four coefficients, u_i through u_l),the coefficients associated with inclusion in each pair of datasets (six coefficients, u_{ij} through u_{kl}), and the coefficients associated with inclusion in three of four datasets (four coefficients, u_{ijk} through u_{jkl}). The coefficients associated with inclusion in multiple datasets are estimates of dependence between the relevant lists. Of course, the saturated model shown above cannot be usefully fit, because it contains no degrees of freedom; at least one coefficient must be dropped in order to create a useful model. More to the point, model selection quickly becomes prohibitively difficult as the number of data systems increases.

For three data systems, there exist "only" (6 choose 1) + (6 choose 2) + … + (6 choose 5) = 81 potential models. For four systems, by contrast, a saturated model has 14 terms and the number of potential models is (14 choose 1) + … + (14 choose 13), or over 16,000 possible models. Clearly, fitting each model and directly inspecting it is not the right solution.

An alternative approach to model selection that is simultaneously more conservative than direct inspection and more tractable in the four-system case is Bayesian Model Averaging (BMA) (Raftery 1995, 1996, Raftery et al. 2005; for application to violence data, see Ball et al. 2007, Lum et al. 2010). BMA employs the same family of log-linear models as the traditional approach outlined above, but accounts for model uncertainty by constructing, in essence, a weighted average across several best-fit models, where weights are determined by Bayesian posterior probabilities. Ball et al. (personal correspondence, 2011) have found that BMA models produce similar estimates to models selected via inspection of the BIC, with the exception that BMA models produce wider confidence intervals because they account for model uncertainty.

My colleagues at the Human Rights Data Analysis Group, Kristian Lum and James Johndrow, have developed a decomposable-graphs (as opposed to log-linear) approach to BMA, which I employ in the homicide estimation that opened chapter 6. The decomposable graphs approach (DGA) was originally described by York and Madigan (1997). Like MSE itself, decomposable graphs were originally employed in natural science applications — in this case, genetics research. As in the MSE case, genetics research includes an extraordinary number of potential covariates but relatively few observations, and consequently understanding the structure of interaction (i.e., non-independence) between models represents a significant challenge. Johndrow, Lum and Ball (2015) have developed an R implementation of BMA with DGA. In a few instances, where even DGA estimates cannot be fit, I turn to the lower-bound estimator provided by Chao (1987), as implemented by Baillargeon and Rivest (2007). Chao's method of moments provides a superior estimator when data are extremely sparse, although—

again, —this is an estimate of a lower bound, *not* (necessarily) a point estimate. I prefer the DGA estimator for all strata where DGA posterior probabilities are neither flat nor multimodal.

Matching Although model selection is a key issue for MSE, arriving at the stage in which one considers model choice and other theoretical issues requires first accessing multiple convenience datasets, rendering them readable and clean, canonicalizing the datasets, and matching cases across all datasets with one another. Of these, matching is typically the largest technical hurdle that must be cleared before estimation can proceed. However, I discuss it only briefly here, because in the Salvadoran case, I have chosen to hand-match across datasets (i.e., to match records via direct inspection).

As noted above, MSE estimates rely on the assumption of perfect matching between datasets; before model selection and estimation can proceed, a reliable and statistically defensible matching system must be developed and employed. Methods for matching range from single hand-matchers to teams of hand-matchers whose results are examined for reliability (cf. Ball 1999; Ball et al 2002, 2003; Silva 2002) to increasingly complex machine-learning processes suitable for hundreds of thousands of records (e.g., Bilenko et al. 2003, Witten and Frank 2005, Sarawagi and Bhamidipaty 2002, Klingner 2007). Virtually every human rights data analysis requires automated matching because of the sheer magnitude of potential matches: examining every potential pair among N records to determine whether it is a match requires the examination of N2/2 potential pairs. Thus, were we to examine every potential match pair, matching the roughly 20,000 documented killings during civil war in El Salvador requires in excess of 200 million observations.

A few statistics on the matching process and results are worth noting here. Across all types of violence and all four datasets, there were approximately 66,000 observed episodes of violence (where an "episode" refers to one act of violence to one victim). Matching proceeded after eliminating from the matching process several thousand unmatchable observations (those with too little information to be reliably matched). The matching algorithm eventually identified approximately 50,000 unique episodes of violence, of which over 45,000 were recorded in only one dataset. The match rate was higher for lethal than for non-lethal violence, although the magnitude of the difference depended on other factors, including the year, the particular non-lethal form of violence, and the geographic region of the country.

Practical issues MSE is the only method by which convenience data can be used as a basis for social scientific inference. However, this validity comes at a high cost. Stratum estimates are derived from individual reports of violence; hence, the number of observations usable in a social scientific analysis (the "N" that typically represents the number of observations in a social scientific analysis, not the unknown population size discussed in the MSE derivations above) decreases from the number of directly observed episodes of violence (about 65,000 in the Salvadoran case) or the number of spatiotemporal observations into which the data are frequently divided (approximately 2400 municipality-years in the Salvadoran case) to the number of strata for which MSE estimates of reasonable quality are available (about 200 in this dissertation). Adding to this difficulty is the fact that many strata estimable using MSE will overlap. For example, in the Salvadoran case, a few of the usable strata include killings in 1983, killings by state forces in 1983, all violence by state forces in 1983, and killings in the department of Chalatenango in 1983.

Rather than viewing this "disappearing N" as a reason to abandon the use of MSE, however, I propose that it be viewed as evidence that traditional sub-national analyses of political violence are only artificially powerful for inferential purposes. This project's theoretical interest in repertoires of violence (as opposed to single forms of violence such as killing or sexual violence) presents important measurement challenges. First, it seems clear that in most cases, while deaths are under-reported, other forms of violence are significantly more under-reported. For example, in the Salvadoran case, while sexual violence was in relatively frequent use by government troops, all four datasets include extremely sparse evidence of sexual violence. UNTC data are by far the least complete in this respect; the Truth Commission's data collection process clearly focused on lethal violence to the exclusion of most other forms. Indeed, approximately 85% of observations in the two UNTC datasets represent episodes of lethal violence (either killings or disappearances). CDHES and El Rescate data contain a larger proportion of non-lethal episodes of violence.

A final important issue related to MSE on non-lethal episodes of violence is: what do these estimates mean? An MSE estimate on lethal violence clearly refers to both one victim and one violation. MSE estimates on torture, however, could refer to episodes of torture or to victims of torture, because victims can (and in many cases do) suffer repeated torture. Moreover, we must ask what is to be counted as an individual episode of torture. Does every torture session in a month-long detention count? Is the detention period the true unit of analysis? In examining hand-matching pairs, I have developed an entirely non-theoretical answer to this question: matching distinguishes confidently between persons, not between episodes of violence. For

example, two records with matching names and locations, whose dates of death differ by a week or a month, are marked as a match with near certainty. In the case of torture, these episodes may or may not refer to the same episode; we can be relatively certain only that they refer to the same individual victim. Thus, in my MSE analyses of non-lethal violence, the unit of observation is the individual victim, regardless of the number of times that victim may have suffered torture (or sexual violence, or any other non-lethal form of violence).

4. Informed consent procedure and semi-structured interview protocol, 2015

Administered by me, in Spanish, at the beginning of every interview. Occasionally, my research assistant would offer clarification. I used the same protocol in 2008-2009, but wording differences are shown in brackets.

Hello, thank you for meeting with me today. My name is Amelia Hoover Green. I am a researcher from Drexel [Yale] University in the United States. This is my research assistant, Erika Murcia, who will be helping with the interview today.

I'm doing research about training and discipline in different armies, and I am focusing on El Salvador. As part of that research, I am interviewing ex-combatants from all of the groups that were involved in the civil war in El Salvador. Today I plan to ask some questions about your experience during the civil war in El Salvador. I will ask some questions about how you joined [the guerrillas/the army/the security forces], some questions about the training you received, and some questions about discipline while you were in [the guerrillas/the army/the security forces]. The interview should take about one to two hours.*

Being interviewed is completely voluntary, and you can stop the interview at any time. Do you want to go ahead with the interview?

[Wait for response.]

It is helpful to have a record of your exact words. I will keep these recordings secure and private. I keep all the recordings in a secure computer file and I do not share them with anyone, even Erika. You do NOT have to consent to have your interview recorded, and you can ask me to stop recording at any time, and I will stop. Do I have your consent to record?

[Wait for response.]

You can refuse to answer any question, or stop the interview at any time. I will do everything possible to keep our conversation confidential. I will never identify you by name in any of my writing, and I won't use any information that would allow someone to guess your identity. I will be taking notes today, but without using names. As soon as I take my notes they are stored in a safe, locked place. This paper [a 4'' by 6'' card with interviewee rights and contact numbers] explains your rights as an interviewee. If you have any problems related to this interview, you can contact me, or my colleague George, who speaks fluent Spanish, or the office at Drexel that protects people involved in research.

Do you have any questions before we begin?

OK. Considering everything that you heard, do I have your consent for this interview?

[After obtaining consents:]

[Note: Parentheses indicate potential follow-up questions; brackets indicate notes to the interviewer.]

1. During the war, which group did you primarily serve with?
   - 1a. [guerrillas] Which group in the FMLN were you in? Were you part of an urban commando unit, a militia, or were you a regular combatant?
   - 1b. [state forces] Were you in the Army? Which part of the army did you serve with? (Which brigade/battalion/detachment?)
   - 1c. [state forces, not in army] Did you serve with the security forces? Which group in particular? (Policia Nacional, Guardia Nacional, Policia Hacienda, other?)
2. Did you ever change forces, for example from one FMLN faction to another, or from the army to a security force?
3. When and where did you enter [group]? When and where did you leave?

4. Where were you posted during your time in [group]?

   ...And after that?...

   [Get a complete list of postings if possible.]

5. How did you come to join [group]?

6. Did you feel you had a choice to join or not to join? Why or why not?

7. [if incomplete answer to 6] Would you describe your recruitment as voluntary, forced, or something in between? (Why?)

8. What was the process to join up? (For example, was it formal, with papers to sign, or informal, where you just show up?)

9. How did you feel about joining at the time?

10. Did you join with friends or people you knew already?

11. After joining, did you receive training?

12. What types of training did you receive? (For example, in the United States Army people learn how to handle weapons, get physical fitness training, and learn certain job skills, like radio or mechanical skills.)

13. How much training (how many days, weeks or months) did you receive before beginning [job/assignment]? (Or did you begin to have assignments immediately?) [for FMLN] (Many of the accounts of the FMLN say that training occurred during combat, so that recruits showed their skills by doing harder and harder assignments. Was your experience similar?)

14. From training, would you say you learned anything about the overall purpose or point of the war? [if yes] What did you learn? Would you say you learned anything about who or what you were fighting for? [for state combatants] (What did you hear about Communism or anti-Communism during your training or afterwards?)

15. From training, would you say you learned anything about who or what you were fighting against? [if yes] What or who was that? Why were you fighting against [that/them]?

16. Can you describe a typical day while you were training?

17. Who did you spend most of your time with during training? (Why?)

18. Who gave you your orders while you were [at camp/ in barracks]? Who gave orders in the field, when you were on an operation?

19. Did people always follow orders, or did it depend on the time or the situation?

20. How did you get along with superior officers? What about other people you knew? (Were there particular officers who were known as really tough, or lenient, or different in any way? Why?)

21. Aside from direct orders, were there rules to follow all the time, or did it depend on the time or the situation?

22. What were the most important rules that you can remember now?

23. How did you know when a rule was important? (Did someone explain why the rule was important?)

24. What could happen to someone who disobeyed an important rule? (Was there a punishment?)

25. Were there rules that people saw as less important?

26. What could happen to someone who disobeyed a rule that was less important?

27. No one is on maneuvers (fighting) all the time. When you weren't on maneuvers (fighting), what was life like? (How did you spend your time?) (Did you have a lot of free time or not so much?)

28. [for FMLN] The FMLN had a rule against alcohol in most of its zones. Did you think this was an important rule?

29. Were there rules for distinguishing between combatants and non-combatants?

30. Were there specific rules about how to treat non-combatants? What were those rules?

31. Did you ever see other people in your group mistreating non-combatants—for example, being violent toward them, or stealing? Can you describe that?

32. Did commanders find out when other people in the group mistreated non-combatants? [If yes:] How did they usually find out, and how did they usually react? Was there usually a punishment?

33. Was there ever a time when discipline really changed? (Were there some situations where certain rules didn't apply, or when discipline got stricter?) Why do you think it changed?

34. Is there anything else you think is important about your recruitment, training, or discipline?

35. Is there anyone else you think I should talk to? (Why?)

5. Informed consent procedure and structured interview (survey) questions

The following material was administered as a two-page paper survey, in Spanish. Frequently, though only at respondents' request, the survey administrator (my research assistant) read both the information/consent script and the individual questions.

Survey About Experiences of Recruitment, Training, and Discipline

This survey is about recruitment, training, and discipline in armed groups that fought in the Salvadoran civil war. "Armed groups" include formal groups like regular and special army units, mobile strategic forces of the FMLN, and state security forces. It also includes informal groups like local paramilitaries, militias, and others.

The survey is ANONYMOUS. We will never ask for your name. If you are uncomfortable responding to any question, you can simply leave it blank. You may also stop taking the survey at any time. The results of the survey will only be known to the researchers. The results will not affect your opportunities for veterans' or any other benefits.

It should take about 15 minutes to complete the survey. [*If reading survey aloud: Do you want to start the survey now?*]

We thank you for your help.

1. How old are you now?

   (Enter a number.)

2. Are you male or female?

   (Enter M or F.)

3. Where do you live now?

   (Enter department and municipality or village.)

4. How old were you when you joined an armed group?

   (Enter a number)

5. In what year did you join?

   (Enter a number)

6. If you don't remember what year you joined, did you join before the beginning of the war, in the beginning of the war, in the middle, or toward the end?

   (Enter pre-war, beginning, middle, or end)

7. Where did you live when you joined?

   (Enter department and municipality or village.)

8. What was the highest grade of school you finished before joining the group?

   (Enter a number)

9. Would you describe your recruitment as forced, voluntary, or something in between?

   (Enter forced, voluntary or in between.)

10. What group or unit did you join, specifically?

    (Enter a number or location of your battalion, detachment militia or other group.)

11. Where was the camp or base of your unit?

    (Enter a department and municipality or village.)

12. How much training did you receive?

    (Choose less than two weeks, two weeks to one month, one to three months, or more than three months.)

13. Did you ever leave El Salvador to receive training?

    (Enter Y or N.)

14. If you left El Salvador for training, where did you go?

    (Enter the country visited.)

15. How were the relations between officers and enlisted soldiers during training?

    (Choose almost always bad, more bad than good, mostly neutral, more good than bad, or almost always good.)

16. Were you ever physically injured during training, as punishment or for another reason?

    (Choose Y or N.)

17. If you were injured in training, what types of injuries did you receive?

    (Mark all that apply: beaten; beaten with an object; kicked; cut; shot; other.)

18. During training, were you ever humiliated, threatened or psychologically injured, as punishment or for another reason?

(Choose Y or N.)

19. Many people report different experiences of punishment during training. Did you experience any of the following things during training?

    (Mark all that apply: verbally abused; threatened with violence; punished in front of others; made to do embarrassing or menial work; made to do humiliating work; forced to dress like a woman; made to do a filthy chore; made to do violence; other.)

20. During training, did you receive political education, for example, a class or lecture about the political reasons for the war?

    (Choose Y or N)

21. If you received political education during the war, what did leaders (officers, commanders, etc.) say about the reasons for the war.

    (Enter a brief open-ended response.)

22. During or after the training period, did you receive books or pamphlets about history or political themes from commanders?

    (Choose Y or N)

23. If you received books or pamphlets, try to recall their titles and what they were about.

    (List titles or other information about books or pamphlets.)

24. How important was discipline in your unit/group?

    (Choose extremely important, important, not very important, or not at all important.)

25. How often did leaders of your group give instruction about how to treat civilians?

    (Choose never, a few times, several times, a few times a week, nearly every day.)

26. During the war, did it seem to you that treating civilians well was a priority for your officers/leaders?

    (Choose Y or N.)

27. Even extremely disciplined groups occasionally make mistakes. Did you ever hear about members of your group disrespecting or mistreating civilians?

    (Choose Y or N.)

28. Did you ever hear about others in your group committing the following forms of violence against civilians?

    (Mark all that apply: verbal abuse; killing domestic animals, stealing food or money, rape or other sexual violence, torture, property destruction, killing, other.)

29. If other combatants behaved badly toward civilians, did your leaders punish them?

    (Choose always, usually, sometimes, usually not, or never.)

30. In your opinion, did displine in your group improve, worsen, or stay about the same as time went on?

    (Choose improve, stay about the same, worsen.)

31. Did you ever change units during the war? For example, did you ever move from regular forces to special forces?

    (Choose Y or N.)

32. If you changed units or groups during the war, what group or unit did you change to?

    (Enter the name of a group or unit.)

33. If you would like to share other information or opinions about recruiting, training, or discipline, please share that information here.

    (Enter short answer.)