From events to data: Politics and the production of government records

Scott J. Cook (Texas A&M University) and David Fortunato (University of California, San Diego)

Issues of measurement have received considerable attention in political science. This is because, as Munck, Møller and Skaaning (2020, 347) argue, "The social sciences. . . are factual sciences, given that they refer to facts about the concrete world. Thus, empirics, and more narrowly measurement. . . are essential parts of social science research." Many discussions of measurement tend to, rightly, focus on the variety of challenges researchers face in producing data (e.g., concept formation, indicator development, etc.). Here we set aside most of these concerns and focus on a narrower issue: measurement error in government data. Government data—records collected and disseminated by governments and government-adjacent international organizations—are widely used in political science research, including election results, crime statistics, trade data, etc. With these data, researchers evade many of the traditional concerns of measurement, as there tends to be broad consensus on the underlying concepts, indicators are based on facts, and the data are already collected. Despite this, we consistently find that even "high quality" government data are often plagued by mismeasurement, which left unaddressed threatens the validity of not only inferences but basic description.

While political scientists, notably Hollyer, Rosendorff and Vreeland (2014), have studied governments' choices to report data to external actors, many continue to (implicitly) assume that government data are accurate, or, that the manipulation of government data is exclusive to autocracies. In the following, we discuss how government data collection and reporting processes create opportunities for systematic errors even in wealthy democracies. We then show that misreporting and underreporting is prevalent in official U.S. data on crime and policing—data collected in a well-resourced democracy, purporting to convey objective facts on highly salient issues—which makes obtaining accurate estimates on events like killings by police extraordinarily challenging. This, we argue, is not because the collection of these data is inherently difficult, but instead a logical result of the political processes that shape whether and how government data are produced.

As such, we feel there is a need for further research into the political economy of government data. Scholars must consider the literal data-generating (or data-production) process separately from the outcome-generating process which tends to be the focus of our theoretical interest. Not only are questions on the data-generating process—government transparency, accountability, and the politics thereof—interesting in and of themselves, but they also have clear implications for any social science research utilizing these data. To aid in this, we outline how researchers can begin to engage questions of data production and discuss the consequences of failing to do so.

Politics and the production of government data

Decisions about the collection and dissemination of data by governments *are* policy choices. As such, they warrant scrutiny by researchers, especially political scientists. To focus our discussion, we concentrate on event data—Schrodt (2012, p. 548) defines an event as a "discrete incident that can be located at a single time (usually precise to a day) and set of actors,"—such as a death, vote, payment, etc. The outcome of interest is the occurrence of the event itself, with the determinants of this informing the *event-generating* process. Conditional on the realization of this event, it is either accurately recorded in government data or not, which constitutes the *data-generating* process. Despite our focus on event data here, the concerns we raise on data quality apply broadly to various other types of government data.

From events to data

While the specific data-generating process for any issue is unique, there are a set of minimal questions that researchers should consider when using government data. First, what is the process through which events are memorialized? Are the events automatically submitted into an archive (e.g., weather indicators, legislative proposals), or, do the events require a party to voluntarily report or document them (e.g., crime, death). Second, who enters the event into the record? Reporting the event may be labor or expertise intensive, creating inequities due to differing resources,

From events to data (cont.)

Cook and Fortunato

or, recorders may manipulate the data due to personal political preferences. Third, why was the event recorded, or, for what reason are the data being collected? Some government data, such as those provided by the National Weather Service, are collected for their own sake, in an effort to transparently provide a public good (supporting commerce and public safety). Yet, other data are collected with clear policy objectives in mind, that is, to accomplish a specific task. For example, data on the number of public school students with special needs, the services provided to them, and the performance of those students are required by the No Child Left Behind Act in order for schools to access federal funds provided by the Individuals with Disabilities Education Act. It seems likely that data collected for their own sake as a public good and data collected in pursuit of funding transfers are prone to different types of error in recording, aggregation, and dissemination. Often it will be useful to map out the process by which events become data.

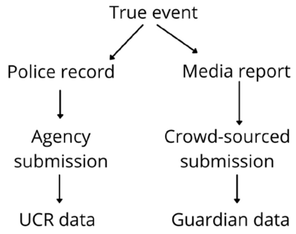

Figure 1: Alternative data-production processes from a single event.

Drawing from recent research by Cook and Fortunato (2022) on killings by police, Figure 1 demonstrates two ways by which a true event may become data: government data, (the FBI's Uniform Crime Reporting [UCR] program) vs. crowd-sourced media reports (The Guardian's The Counted data). Focusing on the UCR data production, we observe that initial reports are filed by individual officers. These reports are aggregated by the agency and submitted to the FBI for inclusion in the UCR database. As such, there are several steps in this process which may cause true events to go unreported in the final UCR data. Individual officers may be tempted to file inaccurate reports (e.g., Egel, Chabria and Garrison, 2017) or not report at all. Individual agencies, then, have discretion on whether to enter events into their UCR submission (or submit to UCR at all). Our research shows that many choose to redact police killings from their reports—for example, the Guardian's The Counted Data verified 71 killings by police in Florida in 2015, yet none are entered into the UCR. Finally, once the data find their way to the UCR, those running the program have additional discretion. For example, the 2016 UCR data, the first released under the Trump cabinet, were conspicuously less detailed than in previous years (Malone and Asher, 2017).

Setting aside incentives for outright manipulation, there is substantial variation in law enforcement agencies' resources for aggregating and submitting data to the UCR; many agencies are quite small, employing few officers and civilian support staff. Between 1995 and 2017, over 25% of agencies had less than three *total* employees, meaning that many police agencies may simply lack the resources to comply with data requests. To illustrate this point, we gather data on UCR participation for all 19,095 state, county, and city police agencies for the 1960-1994 and 1995-2017 periods[1] and regress participation (whether an agency submitted data) on two proxies for agency resources: the total number of agency employees and the size of the population it serves (both rescaled to standard normal).

The results in Table 1 indicate large, positive correlations between our proxies for agency resources and UCR participation. This suggests that we likely have significantly less (and poorer) data on crime in smaller (and poorer) communities.

Table 1: UCR compliance and proxies for agency resources

	1960-1994		1995-2017	
Employees	0.026***		0.016***	
	(0.001)		(0.001)	
Population		0.041***		0.034***
		(0.001)		(0.001)
State FE	x	x	x	x
Year FE	x	x	x	x
Observations	668,325	668,325	452,226	452,226
R^2	0.213	0.217	0.115	0.120

Note: *p<0.1 **p<0.05***p<0.01

These are critical considerations for applied researchers when using these data. Our reading of the extant literature on crime and policing, however, suggests that most prior research using UCR data has not carefully considered these issues. This is particularly troubling given

[1] We separate the periods because they come from different sources. The latter are supplied by the FBI in a standard spreadsheet, the former were parsed from oddly formatted or unformatted text files received as part of our FOIA request.

From events to data (cont.)

Cook and Fortunato

that UCR reports missing data cells as *zeros*. This means that UCR zero counts may indicate that the true number of events was zero, that the agency incorrectly reported a zero, or that the agency failed to report any information on that event-type. Yet, the prevailing method for handling UCR zeros in economics research is to treat them as *true counts*, unless the agency submitted nothing at all that year (e.g., Mello, 2019; Weisburst, 2019). This means that agencies that report the number of officers employed, but no other data, will enter empirical analyses as having zero police killings, zero sexual assaults, zero property crimes, etc.

Consequences of imperfect data

When researchers do not separately consider the outcome- and data-generating processes—accepting the data as (near) perfect memorialization of events—they are of course more likely to draw incorrect inferences. The particular nature of these threats to inference depends on how these data are used and whether the outcome- and data-generating process share common determinants. For example, if the events themselves are the unit-of-analysis, then unreported events would induce a form of sample selection bias. More typically, events are located and aggregated into spatial-temporal units (e.g., state-year), where unreported events instead induce measurement error in the outcome. At best, this will produce attenuation toward the null, however, we cannot safely assume this as many of the same features that cause the outcome also cause variation in reporting rates, risking bias in either direction (Carroll et al., 2006)[2]. For example, Glaeser and Sacerdote (1999) use data from the UCR and the National Crime Victimization Survey (NCVS) to compare crime incidence across municipalities. They find a very large, positive correlation between population and crime (more crimes per resident as population grows), but puzzle over the substantially smaller correlation between crime and population—a correlation that is *negative* when comparing cities of 25,000 and greater—when examining survey responses. Given the UCR reports missing values as zeros and the strong correlation between city size (and agency staffing) and agencies' propensity to report data into the UCR, the more likely reason for this gap is that the survey data (assuming the sample is well-calibrated and representative) are providing a more accurate estimate of crime victimization.

Assuming for the moment that the NCVS provide accurate estimates of crime victimization, what would have to be true in order for the UCR and NCVS to provide effectively identical estimates? Returning to discussion above, 1) all victims must report their victimization; 2) all responding officers must accurately memorialize the incident and file the report (a step Eckhouse, 2021 demonstrates is prone to significant manipulation); 3) the agency must submit all reported crimes into the UCR; 4) the FBI releases all UCR data to the public. Without considering agencies' incentives to manipulate, this process chain allows (at least) four opportunities for attrition—victims may choose not to report, responding officers may make filing errors, agencies may fail to comply with UCR, the FBI may not release all information—but almost no opportunities for over-counts apart for a small number false-reports (which are themselves a crime) in step 1. That is, even if the only error in the process is "random," the net effect is still inherently asymmetric, producing lower estimates of the base rate of crime. Given the relationship between agency resources and reporting, these errors are more likely (in practice, larger undercounts of crime) in smaller or poorer cities, inducing further bias. These errors have implications not only for academic research, but government policy, as policymakers utilize these data unaware of their limitations.

Conclusion

Our aim is to present potential issues in the collection, aggregation, and dissemination of government data. While these data are widely used to study many events of interest (e.g., auto accidents, high school graduation rates, unemployment, etc), too often researchers fail to consider their limitations. All government data have a formally or informally mandated data-generating process that risks error given the nature of the process, the actors involved, and their incentive structure. We illustrate opportunities for significant manipulation and selection bias using the data-generating process underlying the UCR that there is strong positive correlation between UCR compliance and agency resources, and discussing researchers' insufficient consideration of the UCR's data-generating process. Because of this failure, it is our judgement that nearly

[2] This problem is further compounded if both the outcome and input data come from the same source, risking potential "common source" bias (Favero and Bullock, 2015).

every study of crime or policing which naïvely employs UCR data cannot be trusted. How many other fields of study treat similarly imperfect data as unbiased samples?

While we have focused on data from government records, these issues broadly apply to other data sources (e.g., media reports, historical accounts) frequently used in political science research. In peace studies, for example, researchers often use data on political violence (ex. Social Conflict in Africa Database) drawn from news media reports (ex. the *Associated Press*). As a result, events in some countries (e.g., lesser developed states) tend to be underreported (Hendrix and Salehyan, 2015), and, even within a particular country, events in some areas (e.g., capital cities) are more likely to be reported (Weidmann, 2016). The systematic errors in the reporting of these events have consequences identical to those above (as discussed in Cook and Weidmann, 2019). In light of this, many conflict researchers have increasingly used official data (Weidmann, 2015) or conflict archives (Balcells and Sullivan, 2018) when available. In some cases these non-media data may be more reliable, however, it should not be assumed that they will be given our discussion here. Instead, we encourage researchers to scrutinize their data—asking at least how events are recorded, who records them, and for what purpose—regardless of the original source(s).

Beyond the simple recognition of the limitations in one's data, how should researchers proceed? While much of the specifics will have to be addressed elsewhere, in short there are two ways forward. First, where possible, researchers should try to find multiple sources of data on their phenomena of interest, ideally ones in which the preferences and the priorities of the data producers diverge (e.g., contrasting data collected by police departments with data collected by media). Minimally, this will allow researchers to compare findings across alternative data and ensure that any inferences drawn are not source, and therefore *process*, sensitive.

Second, analysts should consider explicitly modeling their uncertainty over data quality. Without additional data, this will typically take the form of sensitivity analysis or bounding, approaches that may be particularly helpful when the shape of the potential bias can be inferred (e.g., Knox, Lowe and Mummolo, 2020). With multiple sources of data, the available options are much richer, as most measurement error models require some type of validation or replication data (Carroll et al., 2006). For example, Cook et al. (2017) demonstrate how with two sources of conflict data, researchers can analyze both the probability of event and report, that is, specify models of both the outcome- and data-generating processes, respectively. We feel that future research in this area is especially worthwhile, as the variety of available data sources and types continues to grow. As such, better understanding how to effectively integrate multiple sources of data to obtain more accurate results is likely to be a fruitful area for research.

References

Balcells, Laia and Christopher M Sullivan. 2018. "New findings from conflict archives: An introduction and methodological framework."

Carroll, Raymond J, David Ruppert, Leonard A Stefanski and Ciprian M Crainiceanu. 2006. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Cook, Scott J, Betsabe Blas, Raymond J Carroll and Samiran Sinha. 2017. "Two wrongs make a right: Addressing underreporting in binary data from multiple sources." *Political Analysis* 25(2):223–240.

Cook, Scott J. and David. Fortunato. 2022. "The Politics of Police Data: State Legislative Capacity and the Transparency of State and Substate Agencies." Working Paper.

Cook, Scott J and Nils B Weidmann. 2019. "Lost in Aggregation: Improving Event Analysis with Report-Level Data." *American Journal of Political Science* 63(1):250–264.

Eckhouse, Laurel. 2021. "Metrics Management and Bureaucratic Accountability: Evidence from Policing." *American Journal of Political Science* in press.

Egel, Benjy, Anita Chabria and Ellen Garrison. 2017. "Hands removed, findings changed: Pathol- ogists say San Joaquin sheriff 'does whatever he feels like doing'." *Sacramento Bee* December 05, 2017 .
URL: *Access*

From events to data(cont.)

Cook and Fortunato

Favero, Nathan and Justin B Bullock. 2015. “How (not) to solve the problem: An evaluation of scholarly responses to common source bias.” *Journal of Public Administration Research and Theory* 25(1):285–308.

Glaeser, Edward L and Bruce Sacerdote. 1999. “Why is there more crime in cities?” *Journal of political economy* 107(S6):S225–S258.

Hendrix, Cullen S and Idean Salehyan. 2015. “No news is good news: Mark and recapture for event data when reporting probabilities are less than one.” *International Interactions* 41(2):392–406.

Hollyer, James R, B Peter Rosendorff and James Raymond Vreeland. 2014. “Measuring transparency.” *Political analysis* 22(4):413–434.

Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. “Administrative Records Mask Racially Biased Policing.” *American Political Science Review* pp. 1–19.

Malone, Clare and Jeff Asher. 2017. “The First FBI Crime Report Issued Under Trump Is Missing A Ton Of Info.” *FiveThirtyEight* October 27, 2017. URL: Access

Mello, Steven. 2019. “More COPS, less crime.” *Journal of Public Economics* 172:174–200.

Munck, Gerardo L., Jørgen Møller and Svend-Erik Skaaning. 2020. Conceptualization and Mea surement: Basic Distinctions and Guidelines. In The *SAGE Handbook of Research Methods in Political Science and International Relations*, ed. Luigi Curini and Robert Franzese. London: SAGE Publications chapter 19, pp. 331–352.

Schrodt, Philip A. 2012. “Precedents, Progress, and Prospects in Political Event Data.” I*nternational Interactions* 38(4):546–569. URL: *https://doi.org/10.1080/03050629.2012.697430*

Weidmann, Nils B. 2015. “On the accuracy of media-based conflict event data.” *Journal of Conflict Resolution* 59(6):1129–1149.

Weidmann, Nils B. 2016. “A closer look at reporting bias in conflict event data.” *American Journal of Political Science* 60(1):206–218.

Weisburst, Emily K. 2019. “Safety in police numbers: Evidence of police effectiveness from federal cops grant applications.” *American Law and Economics Review* 21(1):81–109.