Data Quality for Measuring Political Protest and Government Change

Charles Lewis Taylor  
*Virginia Tech University*

**Abstract**

For analyzing protest and government change, data need to be both reliably measured and valid for the task at hand. Data fundamentalists insist that nothing less than total capture of every event that has taken place is sufficient for analysis. That scenario is, of course, an impossibility. Rather, methods to measure the statistical probability that a sample is appropriate for its targeted universe allow empirical analysis to proceed.

**Keywords:** Reliability, validity, event indicators, political protest, government change

1. Introduction

Analyzing government change and political protest can contribute to understanding causes of conflict, which in turn may lead to its mitigation. For this analysis to be effective, however, reliable empirical data for testing theory are as important as the theoretical frameworks they are to test. Otherwise, we obtain only “grand theories firmly planted in mid-air,” as Karl W. Deutsch used to say. Assuring the quality of data, therefore, is itself a serious component of scholarship that raises the argument above the level of ideology. Measuring political phenomena is fraught with dangers derived from imprecise or false concepts as well as from clerical error. Assumptions and procedures in data coding affect analytical results and are therefore inherently theoretical in nature. Clear-headed thought in developing political measurements is thus required if progress is to be made in political understanding.

Data are judged by their validity and their reliability. Validity refers to cogency or soundness in reason. Are the procedures and rules for coding able to withstand counterarguments? Are they convincing? Do they work for the theoretical purpose at hand? Reliability refers to trustworthiness and accuracy. Are the data dependable? Can they be used with confidence? Are they sufficient for the questions being asked of them? Validity is one among several exercises in theory building. It requires speculative and theoretic contemplation. Reliability, on the other hand, is more a matter of meticulously disciplined work. It involves simply getting it right. If a dataset is to be effective for analysis, it must be constructed both with carefully designed concepts and closely controlled accounting.

2. How Good Do the Data Need to Be?

This question is meaningful only in terms of how the data are to be used. Are they suitable for the purpose to which they will be put? Considerably more detail will be needed for a fine-grained study than for an all-embracing comparison. Examining the exact sequence of...
conflict events during an eight-day riot will require far more complete data than investigating
the differing levels of aggregate violence in a number of societies over a century. Data
fundamentalists find this concept difficult to accept. They assume that sufficient diligence
will produce absolute accuracy and believe that nothing less is acceptable for any analysis.
Not only is there a truth out there, but it is a truth that can be determined with exactitude.
Inevitably, personal perceptions of truth become the standard by which results are judged.
This situation leads either to endless argument or to scholarly nihilism. The alternative that
avoids this dilemma is to analyze human behavior in probabilistic terms. Working with
partial knowledge and estimating the probable likelihood of error is usually the best we can
expect. The development of the statistical approach to scientific analysis was based upon this
understanding.

Charles Brockett missed this vital point in his evaluation of event data reported in the
World Handbook of Political and Social Indicators: Third Edition. He claimed to discover
“gross and systematic errors” in the data because he found reports of some events that were
not included. He asserted that analyses employing this source were flawed due to their use
of “meaningless data” to measure mass political violence. Although he made these charges
on the basis of very short-term observations (over scattered years in the 1970s) of one data
series (political deaths) for a very small number of cases (five Central American countries),
he reified his figures into a benchmark of real truth against which to measure the entire 30
years, 139 countries, and 18 event series of the World Handbook. “A prodigious undertaking,
providing annual events data for 1948-77, it is not good enough,” he concluded.¹ Good
enough for what he did not address.

In addition to ignoring the fact that such a tiny, non-random sample is unrepresentative of
the whole population of events, Brockett seems to have misunderstood the general nature of
statistical analysis. In their devastating reply, Dixon, Muller, and Seligson made the point that
documented evidence of underreporting does not necessarily imply a precise accounting.
Brockett’s investigation of deaths in Central America is a case in point. While he does
supply convincing evidence of underreporting of deaths, he fails to provide enough detailed
information to permit an accurate correction.²

Dixon et al. used a censoring technique that identified moderate rather than the excessive
levels of violence claimed by Brockett.³ Because underreported cases could be reasonably
expected to rank above the threshold that they determined, they set the score for these cases
equal to that threshold. This technique allowed observing how the underreported cases
affected their analysis. When they compared the effects of the adjusted data to those of the
unadjusted data in a model predicting violence from land inequality, the estimates changed
slightly and the overall fit declined marginally, but the general pattern remained pretty
much the same. Making allowances for error and employing the best data available did not
overturn the results. Dixon, Muller, and Seligson noted that “blanket charges of inaccuracy
without supporting evidence can do little to advance inquiry” and that evidence of significant

¹ Charles D. Brockett, “Measuring Political Violence and Land Inequality in Central America,” American Political Science
² William J. Dixon, Edward N. Muller, and Mitchell A. Seligson, “Inequality and Political Violence Revisited: Response,”
³ It would have been possible, they say, to eliminate cases plagued by documented irregularities, but this would have produced
serious distortions since irregularities are located precisely in the times and places where disorder is known to be frequent. The
optimal solution, of course, would have been to correct inaccuracies with better data, but that is easier said than done.
underreporting “is better used as a remedy than an indictment.”

There are no methods by which to obtain a metric of everything that has happened. Historians have of course known this for centuries. There are no perfect scholars, and even if there were, no complete sources for political events exist. No scribe could ever record them all. Events happen more quickly and in more places than can be accommodated by available clerks. Hence, no dataset can be assured of containing every single event of conflict within its scope, whether it be a quantitative assessment for a couple of years of a country’s history taken from police blotters and local newspapers or a carefully compiled descriptive report drawn from several separate historical archives.

3. So What Do We Do?

A dataset needs to be characterized not so much by completeness as by comparability across time and space. Wooley points out that any dataset of political events is in reality a sample drawn from the universe of actual events that occurred, and that no absolute benchmark exists by which the quality of any particular sample can be judged. If we follow the logic of sampling theory, of course the more densely packed the dataset, the greater its likelihood of representing the actual patterns that exist in the total event population. Nevertheless, there is always the overwhelming probability that more events occurred than were reported, however large the sample.

The first requirement for probabilistic analysis is the reasonable assurance that the data sample is unbiased. A sample may or may not be random with regard to the various criteria that are important to the investigation at hand. Moreover, it will be drawn from an unknown population. A careful examination is required of the sources as well as the protocol by which the data are to be coded or otherwise manipulated. Analytical interest, of course, is in the patterns of events, the relationships among these patterns, and the connections between them and other relevant variables. Are the data a sufficient representation of the underlying population of events for testing probabilistic models that explain these patterns and relationships? The answer depends upon the requirements of the particular model. Dixon, Muller, and Seligson found no difference in relationship when they modified their data, but not all adjustments and retesting would allow statistical results to stand. It is the responsibility of the analyst to assess empirically the effects that data quality will have upon the analysis.

The questions of validity and reliability are of course related to the quality of the sources. How comprehensive is the source? What biases does it contain? Are events from reports recognized (whether by human coders or by automated coding) in terms relevant to theoretical concerns? Answering the last question requires developing base codes consistent with the theoretical framework: What needs to be measured and what is the logic for doing so? Finally, for large datasets that are most efficiently collected by automatically coding digital data, reliability has to do with the convergence of machine and human coders in assigning events to the proper event type and in identifying the associated attributes of these events.

Many of the early cross-national political event datasets were based upon print media. Crews of individual coders worked together to create and refine coding rules that could be

---

4 Dixon, Muller, and Seligson, “Inequality and Political Violence,” 990.
applied to journalists’ reports of activities related to political protest and government change. The participants then checked inter-coder reliability by means of tests on sample material. Such a project was, of course, labor intensive and made even more so by the use of multiple sources, including sources in local languages. Adding to the required work, several scholars have sought local and international news sources. These, they argue, provide additional perspectives on political processes. Multiple sources increase the problem of dealing with duplicate reports of the same event, however. Simply adding event reports wherever found does not alone solve the comparability problem, but sources containing more event reports do offer the opportunity of more comparable reporting. Althaus, Edy, and Phalen made the case for using wire services rather than newspapers. Newspapers, they said, have a “news hole” of a specific size and the events reported have to fit into this space. Wire services have no news hole. Their continuous coverage can more easily expand or contract to cover what is actually happening. Wire services therefore include far more news than newspapers can afford to print.

Reuters News Agency was singled out as a particularly good source that covers much that is missed by other news agencies. Althaus et al. argued that the multiple sources available through Reuters Textline Service in Nexis-Lexis do reduce bias and generate more density. In a test, they found Reuters Textline dwarfed several putative benchmarks in Northern Ireland, Poland, and the United States in the number of protest events that it covered. They found that the Radio Free Europe chronology included only 31.2 per cent of the events found in the Reuters’ sources; Uncensored Poland News (Solidarity Newsletter) covered 3.7 per cent; the Bew and Gillespie Northern Ireland chronology covered 10.9 per cent; and the New York Times covered 11.4 per cent. Almeida and Lichbach found that internet sources provided much more complete coverage than the New York Times, Global Newsbank, and the Seattle Times.

Rucht and Ohlemacher rejected wire services on the basis of their then-inaccessibility and of the flood of information that would have to be screened, but over time, various schemes have been developed for machine-aided coding and eventually for direct computer coding. Significant among these are systems developed by Philip Schrodt and by Virtual Research Associates (VRA). Eventually, agreement between human and machine coding has begun to approach 90 per cent, which is at a level about equal to the best results among trained human coders.

A major problem for human coders is the “black box” of the coder’s mind. It is difficult...
to account for exactly how and why any single individual may respond to an event report. Coding manuals, coder training, and inter-coder reliability checks reduce the problem of random reactions, but coding errors do occur and are difficult to identify. In contrast, automated coding is highly consistent and thereby transparent in terms of error patterns. It is also flexible in that once errors are identified, codes can be altered and data rerun, creating a new dataset. In effect, automation replaces numerous moderately and low-skilled human coders with a few highly skilled programmers and protocol developers.

The Virtual Research Associates Reader is one project involved in the automated coding approach that makes use of Reuters Business Briefs. Participants include Doug Bond, Joe Bond, and Charles Taylor. Current examination of the validity of these data as dynamic indicators of intrastate conflict processes is encouraging. To be sure, media selection and misrepresentation present difficulties for conflict analysis, and although the source of the data is imperfect, it does provide a valid trace of conflict events in the aggregate. The question is not whether international news sources literally describe all events that occur but whether indices drawn from them represent cross-national patterns. The VRA project presents the results of a random sampling approach to assessing coding reliability for 50 conflict event types anywhere in the world.

Given the limitation of resources in the real world, there is always a trade-off between intricate measurements of narrower scope that usually produce a larger number of event reports for any specific time and place, and broader studies that allow greater comparative analysis. The broader the scope in time or space, of course, the greater the possibility for overlooking events of interest because the resources that can be devoted to each detail are relatively fewer. New developments, however, continue to mitigate even this problem.

Further Reading


Bibliography


