Erschienen in: American Journal of Political Science : AJPS ; 63 (2019), 1. - S. 250-264
https://dx.doi.org/10.1111/ajps.12398

Lost in Aggregation: Improving Event Analysis with Report-Level Data

Scott J. Cook Texas A&M University
Nils B. Weidmann University of Konstanz

Abstract: *Most measures of social conflict processes are derived from primary and secondary source reports. In many cases, reports are used to create event-level data sets by aggregating information from multiple, and often conflicting, reports to single event observations. We argue that this pre-aggregation is less innocuous than it seems, costing applied researchers opportunities for improved inference. First, researchers cannot evaluate the consequences of different methods of report aggregation. Second, aggregation discards report-level information (i.e., variation across reports) that is useful in addressing measurement error inherent in event data. Therefore, we advocate that data should be supplied and analyzed at the report level. We demonstrate the consequences of using aggregated event data as a predictor or outcome variable, and how analysis can be improved using report-level information directly. These gains are demonstrated with simulated-data experiments and in the analysis of real-world data, using the newly available Mass Mobilization in Autocracies Database (MMAD).*

Over the last decade, the empirical study of conflict and contention has made substantial progress in understanding the causes and consequences of political violence. Many of these gains are due to the increased availability of more fine-grained, event-level data. Documenting political events with precise spatial and temporal coordinates, these data make it possible to study political contention (e.g., protests, riots, strikes) at unprecedented levels of analytical resolution. In conflict research, event data on political violence have allowed researchers to study, for example, how poverty at the local level affects the occurrence of civil war violence (Hegre, Østby, and Raleigh 2009) or how violence diffuses over space and time (Schutte and Weidmann 2011).

While vital for advancing research in this field, obtaining accurate data on political events remains a major challenge. Event data are drawn from a variety of primary and secondary sources, including nongovernmental organization reports, survey interviews, and government records (e.g., police records in Sullivan 2016 and military databases in Weidmann 2016). However, the most widely used source for event data is reports from media outlets, as these cover multiple cases/countries over many years, thereby enabling systematic large-N analysis. In constructing these data sets, news reports are screened for coverage of relevant events (e.g., political protest), and the pertinent event characteristics (e.g., date, location, actors, number of participants) are extracted.

Although they often constitute the best or only data available, secondary-source reports, and media reports in particular, are imperfect for several reasons (Aronson, Fischhoff, and Seybolt 2013; Earl et al. 2004). First, there is often selectivity in reporting, in which a particular source has incomplete coverage of the population of events. Media sources, in particular, often cover some events with a higher probability than others, depending on the location, type of event, or its severity (Hendrix and Salehyan 2015; Weidmann 2016). Second, even when we observe a report, there remain questions about its veracity (Baum and Zhukov 2015; Weidmann 2015)—that is,

Scott J. Cook is Assistant Professor, Department of Political Science, Texas A&M University, College Station, TX 77843 (sjcook@tamu.edu). Nils B. Weidmann is Professor, Department of Politics and Public Administration, University of Konstanz, 78457 Konstanz, Germany (nils.weidmann@uni-konstanz.de).

Thanks to Benjamin Bagozzi, Timm Betz, Brian Lai, and Samantha Zuhlke for their helpful comments, and to Espen Geelmuyden Rød and Sebastian Hellmeier for their help in creating the MMAD data. All remaining errors are our own.

is the information contained in a report reasonably accurate? Given logistical constraints in reporting and audience considerations in framing, there are often reasons to believe such reports do not perfectly capture the events on the ground.

To help guard against this, most event data sets draw from multiple sources of reporting; that is, they "triangulate" the data in hopes of providing broader coverage and offsetting possible reporting biases (Earl et al. 2004). As a consequence of using multiple sources, event coders—whether human or computer-based—frequently encounter different reports about the same event. These reports are then combined, or aggregated, into a single event observation with a given set of event characteristics. The end result is a list of event-level observations that is then provided to researchers. We argue that this practice of pre-aggregating the data leads to at least two problems. First, pre-aggregation necessarily entails a loss of information. We do not know, for example, whether several reports agreed on the number of protesters (making us more confident in the summary estimate), or whether the report values diverged widely (leading us to treat the summary estimate with some caution). Second, absent clear and transparent coding rules for how to deal with multiple, possibly conflicting reports, we are faced with an opaque and, by implication, less reliable coding process.

This, unfortunately, is common practice in the discipline. According to our review of 11 cross-national, media-based event data sets on political protest and violence, the vast majority (nine) rely on multiple sources (e.g., different news agencies).[1] Some data sets provide the number of reports an event relies on, as well as the underlying sources of these reports. However, none provide detailed instructions on how different reports are aggregated into a single event observation.[2] Even in the rare instances in which data are made available at the report level (e.g., in the Phoenix Event Data), there are no recommendations on how to analyze these reports directly or adjudicate between conflicting reports. More generally, recommendations on best practices when coding and analyzing media-based event data are also silent on the aggregation problem (Salehyan 2015; Schrodt 2012).[3]

In this article, we present a simple and intuitive approach for handling multiple reports that avoids the abovementioned drawbacks from pre-aggregation. Rather than collecting and distributing data on *events*, we recommend that researchers record, disseminate, and analyze data at the level of individual *reports*. Report-level data includes all the variables one would typically extract for an event, but provides separate information for each report covering said event. This report-level approach has a number of advantages. First, it more directly parallels the initial data collection, which occurs at the level of reports and not events. Second, it resolves the lack of transparency over aggregation currently found in event-data coding. With a report-level data set, users are given access to all of the information available and can subsequently adopt and compare different aggregation rules. A third less obvious advantage is that researchers can achieve statistical gains from analyzing report-level, rather than event-level, data.

Specifically, report-level data make it possible for researchers to directly address the measurement issues that arise with media-generated data. In short, multiple reports can serve as replicate data on the event, allowing researchers to undertake a variety of measurement modeling strategies. Where concerns are primarily about selectivity, researchers can leverage multiple reports to undertake multiple systems estimation (or related mark-recapture approaches), using different incomplete lists of events and their partial overlaps to generate a tally of the complete population (Ball et al. 2003). Instead, where concerns are primarily about veracity, we demonstrate how the variance across reports can be used to explicitly introduce uncertainty over the event-level estimates, improving the quality of our inferences.[4] Informally, we can exploit our understanding that some of our event estimates are likely better (those with many complementary reports) than others (those with few or diverging reports). In the following, we show how doing so achieves gains over naïve aggregation strategies using multilevel models (where events are outcomes) and simulation extrapolation (where events are predictors).

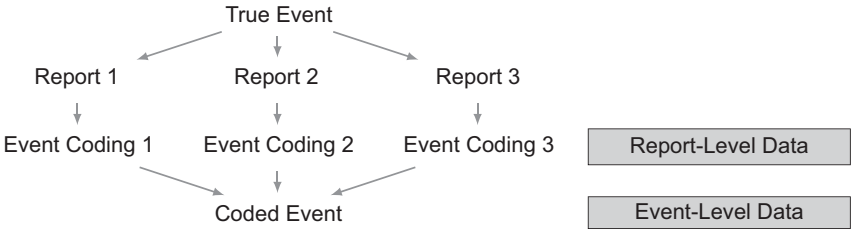
[1]Details on our review of event-level data sets are available in Appendix B in the supporting information (SI).

[2]A related, though distinct, issue is multiple reports of the exact *same* information, such as a news agency report's being published in several outlets. Since these additional reports provide no unique information, they should be purged, via de-duplication, to isolate only independent reports. In our discussion, we assume that reports have been de-duplicated, but we elaborate on this issue in SI Appendix A.

[3]What Schrodt (2012) refers to as the "aggregation problem" is the problem of aggregating events to particular spatial-temporal units of observation; hence, it is unrelated to the problem we identify here.

[4]We privilege field-specific language, yet the measurement concerns discussed find direct parallels in latent variable measurement modeling. What we call "selectivity concerns" relates to the misclassification of latent categorical variables, and what we call "veracity relates" to the reliability of latent continuous variables.

FIGURE 1 One Event Generates Multiple Reports



Note: Each report is coded separately before the reports are aggregated into a single coded event.

Although we discuss the aggregation problem in the context of protest event data and illustrate it with a human-coded event data set from news reports, the applicability of our proposed solution is much wider. First, the strategies we discuss can incorporate and be applied to other sources of data as well. For example, Balcells and Sullivan (2018) advocate data transparency in conflict research using archival (i.e., primary source) data so that researchers will be able to utilize multi-source methods in the future. Second, the aggregation problem applies beyond conflict studies, as data collections in international and comparative political economy also aggregate multiple reports (e.g., Mosley 2011). With the advent of digital publishing and computer-based text processing, the number of sources (and, therefore, reports) available to researchers is likely to increase significantly in the future (Schrodt 2012). For this reason, event coders need transparent and standardized approaches for dealing with aggregation.

In the remainder of this article, we describe the aggregation problem in greater detail and introduce our solution: Collect and analyze report-level data. We then illustrate the advantages of report-level over event-level data in two types of analysis: first, by using report-level data as an outcome variable and second, as a predictor variable. For each of these scenarios, we first use Monte Carlo simulation to analyze how different aggregations and estimators affect performance of the statistical model. We then present the gains from report-level analysis using real data from the Mass Mobilization in Autocracies Database (Weidmann and Geelmuyden Rød 2015), a new report-level data set on political protest.

The Aggregation Problem

Before demonstrating the consequences of analyzing aggregated data, we describe a representative data-collection process in its entirety. Figure 1 shows a simple example. An actual event occurs (top), which is then covered in one or more news reports (Reports 1–3 in the figure). Next, some set of these reports (ideally all) is retrieved and processed by (human or machine) coders. Weidmann and Geelmuyden Rød (2015) refer to this as the *article selection step*, in which the challenge is to include all articles covering the events of interest (e.g., articles on coups, protests, etc.). After collecting relevant articles, the *information extraction step* follows, in which coders (again, human or machine) extract the desired characteristics of an event (e.g., time, location, type, participation level) from each of the reports (Event Codings 1–3 in the figure). This is what we call "report-level data." Finally, in the *aggregation step*, the event codings are synthesized into a single, final coded event. Most widely used event data are presented to researchers *after* this aggregation step, as a final list of events, that is, event-level data.

Although the possibility for errors during reporting (Baum and Zhukov 2015; Weidmann 2015, 2016) and information extraction (Schrodt and Gerner 1994) has received considerable attention, little notice has been paid to the possibility that errors often arise during the aggregation step. Why is the aggregation of multiple reports potentially problematic? Aggregation forces researchers to adjudicate between conflicting accounts, often without sufficient information to accurately render a decision. Clearly, this is not a simple problem. Consider the example in Table 1, taken from the new Mass Mobilization in Autocracies Database (MMAD; Weidmann and Geelmuyden Rød forthcoming). The table shows report-level codings for a single protest incident in the city of Osh, Kyrgyzstan, on March 21, 2005. Article selection returned nine articles from three sources (Associated Press [AP], Agence France-Presse [AFP], and BBC Monitoring [BBC]) on this event, with more than half coming from BBC Monitoring (because of its inclusion of local sources). These articles were separately coded to extract

Table 1 Event Codings for Protest in Osh (Kyrgyzstan) on March 21, 2005 (from Weidmann and Geelmuyden Rød 2015)

Number of Participants	Security Forces Engagement	Source
hundreds	N/A	AFP
2,000	present	AP
1,000	N/A	AP
N/A	present	AP
N/A	present	BBC
1,000	not present	BBC
several thousand	N/A	BBC
3,000	physical intervention	BBC
200	N/A	BBC

Note: N/A means that no information about the respective event characteristic was found in the news report.

information pertaining to the event, such as the number of participants and the level of security force engagement (a categorical variable ranging from *no presence* to *lethal violence*).

A visual inspection of Table 1 demonstrates the practical difficulties faced when combining different reports: How many participants were there? Was there security force engagement? How certain are we about each of the reports? While there appears to be some agreement that the number of protesters was in the low thousands, others estimate a much lower number (several hundred), and others provide no estimate at all. There is even more disagreement when it comes to the level of security force engagement. Four reports contain no information on whether security forces were present—is this missing data or does the absence of information imply no security forces were present? Five reports explicitly mention security forces: One states that forces were present and used physical violence, three state that forces were present but did not use force, and one explicitly reports that security forces were *not* present. Given the lack of correspondence, how do we effectively aggregate this information? We argue any attempt to reduce these reports to single, summary values necessarily risks error.

Although these types of aggregation conflicts are routine in event coding, it is often unclear how they are resolved. As mentioned above, our analysis of existing event data sets revealed that few have (or publicly detail) explicit and consistent procedures for aggregation. Yet, several strategies seem to be the most common. First, some use a single source of reporting (e.g., *The New York Times*). While this does evade the aggregation problem, it does so at the cost of possible data missingness and source bias.[5] Second, coders often seem to employ an informal mental measurement model, using their case knowledge and expertise to classify certain sources or reports as more or less trustworthy. That is, the coder selects the report he or she believes to be the most accurate from the set of available reports. If the researcher is usually correct, there are some clear advantages to this strategy in terms of accuracy. However, this strategy is difficult to replicate, and, more problematically, its accuracy is impossible to externally verify given the lack of gold standard data. Finally, researchers sometimes calculate a statistical summary of the report values—often a mean, mode, or maximum—thereby producing a single estimate of the true value. These statistical strategies are more transparent and can leverage information from each of the underlying reports; still, they restrict the information available to researchers.

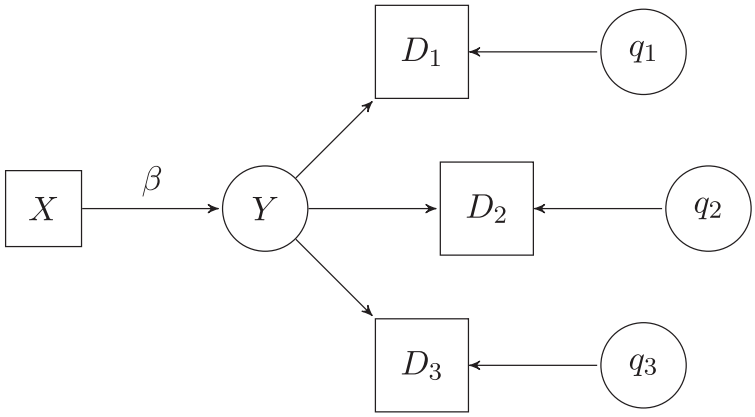
The information on the event is actually produced in the initial report-level coding prior to aggregation, what we call "report-level data." These data contain all of the information about an event that is extracted from reports. From report-level data, one can easily reproduce any of the event-level aggregates—that is, after all, how they are generated. However, the reverse is not true. Information contained in the reports is necessarily lost when data are only provided at the event level. Producing only a final list of coded events hardwires a single aggregation method into the data, eliminating the opportunity to understand, assess, and possibly change it. Furthermore, researchers using event-level data do not acquire information about the variance across report estimates that is *only* available when we have the information at the report level. Not only is the loss of information unnecessary, but as we demonstrate in the following sections, it also has practical implications for our statistical analysis.

Report-Level Data and Outcome Error

Data on many outcomes of interest in comparative politics and international relations (e.g., protests, strikes, boycotts, riots) are not observed or measured by the researcher, but instead derived from media reports (Salehyan 2015). Even though these reports are likely to be imperfect, they often provide the best (or only) available information on these political processes. We visualize the data-generating process in Figure 2: Lacking direct data on the outcome, Y, researchers rely on one or more

[5] A variation on this strategy we encountered was researchers primarily using a single source, but occasionally augmenting it with other reports in an ad hoc fashion.

FIGURE 2 Report-Level Data and Outcome Error



Note: Researchers do not directly observe the outcome of interest Y, but several reports D of the outcome with error q. Square nodes are observed, and circular nodes are unobserved.

reports, D, as proxies.[6] As discussed in the previous section, event data producers employ a variety of strategies to obtain estimates of Y from D. Unless these strategies recover values that closely approximate Y, extraneous error is introduced into subsequent regression models.

Although researchers often ignore, downplay, or fail to address measurement error in the outcome, it can do violence to parameter estimates of interest (Buonaccorsi 1996).[7] The exact consequences of outcome error depend both on features of the model and error process: (1) whether the outcome of interest is continuous or discrete, and (2) whether the error is random (nondifferential) or systematic (differential), that is, whether the outcome error related to the predictors is independent of the outcome itself.[8] Different combinations of these two conditions affect estimates in distinct ways, as summarized in Table 2. In the case most familiar to researchers—a linear model with random measurement error in the outcome—the consequence is greater unexplained error variance, producing less efficient effect estimates and increasing the risk of false negatives in null hypothesis significance testing. In the other cases, measurement error in the outcome also risks bias—either due to misclassification in a discrete-outcome model, or covariance between the measurement error and the predictors. In sum, measurement error in the outcome produces inefficiency at best, and often bias as well.

TABLE 2 Consequences of Error in the Outcome on Coefficient Estimates

	Random	Systematic
Continuous	Inefficiency	Bias (+/−)
Discrete	Bias (−)	Bias (+/−)

These concerns are especially salient in models of event data. First, media reports are widely known to be error prone, as they "may be biased both by political pressure to suppress some reports and by commercial pressure to gain audience market share," in addition to material and logistical constraints on coverage (Aronson, Fischhoff, and Seybolt 2013, 287). As a consequence, media reports can under- or misreport event occurrence, producing misclassification in discrete outcomes (Cook et al. 2017; Hendrix and Salehyan 2015), nonrandom sample selection (Hendrix and Salehyan 2017), or both. While increasing the number of sources (i.e., triangulation) decreases the severity of these concerns—providing broader coverage and negating single-source biases—it does not ensure an exhaustive sample (Earl et al. 2004). Furthermore, drawing on multiple sources of information (including nonmedia sources) raises questions about how to combine and adjudicate between conflict accounts. Naïve aggregation choices (as discussed in the previous section) can induce additional error. Whereas with discrete outcomes, privileging one account over others risks

[6]We follow Buonaccorsi's (2010) notation, with D representing a mismeasured outcome, Y, and W representing a mismeasured predictor, X.

[7]As is standard, our discussion focuses on classical measurement error rather than Berkson errors.

[8]Here, we focus on additive measurement error, yet solutions to multiplicative error are similar to what we discuss: Specify a model of the error process and find enough additional data from which the parameters can be identified.

(additional) misclassification, with continuous outcomes this can result in heteroskedasticity given that some event estimates are drawn from more or better sources.

Not only does aggregation risk greater error, but it also limits the ability of researchers to pursue principled solutions to measurement error. Most solutions to measurement error require finding better or auxiliary data (e.g., replicates, instruments), which are rarely available to researchers using observational data.[9] However, with multiple-source event data, additional data often do exist: the reports. The key insight is to recognize that reports themselves are individual measures of the outcome of interest, with each report providing a noisy estimate of the true event. Moreover, these data are quite easy to collect, as event-data producers already possess the reports. Making them available in this form offers more options for improved analysis than is possible with aggregated event data.

There are several ways in which report-level data can be exploited, depending on the research question. With discrete latent outcomes, existing work has already shown how multiple data sources can be used to address selective reporting (i.e., underreporting) of events (Seybolt, Aronson, and Fischhoff 2013). Specifically, researchers have demonstrated that these methods can recover accurate estimates of population frequencies from several incomplete data sources (Ball et al. 2003; Hendrix and Salehyan 2015). Related work by Cook et al. (2017) has demonstrated that multiple sources of data allow researchers to model both the report-generating and event-generating processes simultaneously in a joint likelihood. However, the application of these methods is necessarily restricted to those areas in which multiple data sources are already available. Making the provision of report-level data the norm, rather than the exception, would facilitate the broader use and refinement of such approaches.[10]

In addition, concerns of event data veracity or reliability can also be better addressed with multiple reports. Consider the familiar linear additive model, where we are often told that random error in the outcome is relatively benign in that it "only" affects the precision of estimates. With knowledge on the number of reports per event—as provided in some event databases—researchers can account for nonconstant error variance by estimating models via weighted least squares (WLS), giving additional weight to those observations with more reports. With full report-level data, researchers can do even better. Report-level data can be analyzed directly in a hierarchical linear model (HLM), explicitly modeling the variance within events (across reports). While this approach is familiar to researchers with panel or mixed-level data (e.g., Steenbergen and Jones 2002), few have recognized its potential with event data. The application of HLMs, or multilevel models more generally, to this context is straightforward given the natural hierarchy in these data: Reports are nested in events and sources.[11] HLMs are able to include all of the given information, with event- and source-level intercepts ensuring that the error structure for the report-level analysis is not heteroskedastic due to repeated observations of the same event or repeated event reports from the same source.

Moreover, these source and event intercepts provide additional insights into the data-generating process. These tell us about the extent of the source- and event-level variance, which may itself be of theoretical interest. As noted by Steenbergen and Jones (2002, 223):

> It is useful to note that clustering and nonconstant variance are often more than statistical nuisances. In many applications, these data features are of substantive interest. For example, what does dependence among the observations imply substantively about the problem under study?

In the case of media-generated data, HLMs allow us to better understand possible source effects. Do some outlets tend to consistently under- or overestimate features of the event? Do some places or times tend to receive fewer or more error-prone coverage? Not only does this help us to better understand and frame analysis, but it also allows interesting theoretical questions about the nature of error in reporting to be addressed. In sum, they help us to resolve the problem of reliability in media reports of events. Yet, these additional insights are only possible with report-level information.

Simulations

To demonstrate that researchers can make better inferences using report-level outcome data—thereby improving their understanding of the political event itself—we undertake a series of Monte Carlo experiments. Data on $N = 1{,}000$ events are generated as

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{1}$$

[9] Gains can also be realized by supplying extra-empirical information (e.g., Bayesian priors), yet data-based solutions are often preferable.

[10] For example, the Cook et al. (2017) two-source estimator can easily be expanded to M-sources as such data become available, enabling researchers to explicitly model correlations across sources.

[11] This is in addition to familiar levels of space (e.g., countries) and time (e.g., years), which are also easily accommodated.

where Y is a true event characteristic, with variation explained by predictor $X \sim \mathcal{U}(0, 1)$ and modeling error $\epsilon \sim \mathcal{N}(0, 1)$.[12] The parameters in the event-level model that researchers are interested in are held fixed at $\beta_0 = 2$ and $\beta_1 = 3$. As in reality, researchers do not possess Y itself, but instead a set of reports on these events. To mirror this, three reports per event D_r from different sources s are generated as

$$D_{r,s} = Y + u_{r,s}, \quad \text{where} \quad u_{r,s} = q_r + v_s, \qquad (2)$$

where $q_r \sim \mathcal{N}(0, \sigma)$ and $v_s \sim \mathcal{N}(0, \tau)$. Modifying σ affects the variation in estimates across reports of the same event, whereas modifying τ allows for common-source effects across events (e.g., if one source regularly underestimates the size of events). In the results reported below, we hold source-level error τ fixed at 0.1, focusing on report-level error by allowing σ to vary from 0.1 to 3.0 in increments of 0.1, with 500 trials evaluated at each setting. In addition, we vary the number of reports per event that are included to reflect real-world variation in reporting effort. Inclusion in the estimation sample is determined as

$$D_{r,s} = \begin{cases} D_{r,s}, & \text{if } Z \geq -0.44, \text{ where } Z \sim \mathcal{N}(0, 1) \\ \emptyset, & \text{otherwise.} \end{cases} \qquad (3)$$

That is, we take draws from a standard normal for each report, eliminating those where $Z < -0.44$. As a consequence, each source "reports" on roughly two-thirds of the events in our simulations. Using these simulated data, we evaluate the performance of several estimation strategies. Our preferred strategy is to use the report-level data directly in an HLM with random intercepts at the event and source level. In this framework, we regress the reports $D_{r,s}$ on X. We compare this to several conventional aggregation and estimation strategies that broadly fall into two categories: (1) those using a single report estimate and (2) those drawing from multiple reports.

Of those that use a single report value, we evaluate two strategies. First, we consider a scenario (*Single Source*) under which the researcher only uses data from a single source of information (i.e., *The New York Times*). We expect this strategy to perform poorly, as it is consistently biased from any source effects. Our second single report strategy (*Best Report*) allows the researcher to make informed choices on which report to use. This parallels data collection efforts in which researchers have case-, source-, or issue-specific knowledge about the data. Specifically, we determine the best report by calculating the minimum absolute distance between the reports $D_{r,s}$ and true Y. We then vary the proportion of times the researcher selects the single best report, which reflects different levels of accuracy researchers may achieve in adjudicating between reports. While questions of this strategy's transparency and external validity remain when applied to real-world data, we expect it should do quite well if the researcher is sufficiently accurate. Both *Single Source* and *Best Report* are estimated via ordinary least squares (OLS), with the results given in Figure 3.

The results, reported in root mean squared error, are consistent with our expectations. At very low levels of report error variance, there is little difference between the various models. This is because each report provides similar information, minimizing the potential gains from using either the best report or leveraging information from each of the reports. As we increase the report error, however, a different story emerges. While all of the models perform worse under greater report error variance, the decreased performance of *Single Source* is the most dramatic. Moreover, *Single Source* is strictly dominated by the other models considered, performing the worst at every level of report error. The best-performing model is *Best Report*, in which the hypothetical researcher selects the best report with perfect accuracy. This is obviously a heroic and untestable assumption with real-world data. When we relax it by allowing for some error in the report selection, the performance of *Best Report* degrades. In Figure 3, we show that when researchers fail to identify the single best report in 50% of the cases, the performance of *Best Report* falls dramatically.

Our preferred strategy, *All Reports: HLM*, returns what could be seen as mixed results. Although it consistently outperforms both *Single Source* and imperfect report selection, *Best Report (50%)*, it is always bested by perfect report selection, *Best Report (100%)*. Therefore, answering the question of the best strategy to use with real-world data hinges on the degree of accuracy one is willing to *assume* for report selection. We emphasize this is an assumption because with real-world data, we never know if we are using the best report. In addition, with event-level data, a researcher cannot even test whether different report selection or weighting strategies would have produced different results. At worst, these results provide support for publishing report-level data *in addition* to event-level data. This would allow for less assumption-dependent hypothesis testing, as researchers could evaluate both the pre-aggregated data and the report-level data.[13]

Continuing on with our simulated experiments, we also compare HLM to a second category of strategies that

[12] Technically, Y, X, and ϵ are indexed by i, which identifies the event. We suppress the indexation to simplify the presentation in Equation (2).

[13] In SI Appendix A, we discuss what researchers should do if results differ.

FIGURE 3 Mean Squared Error for $\hat{\beta}_1$ across Different Levels of Report-Level Variance

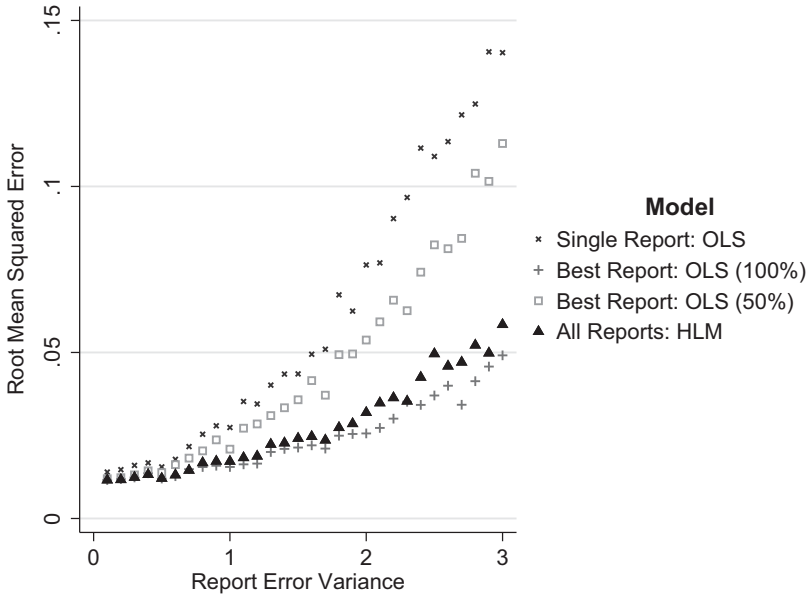
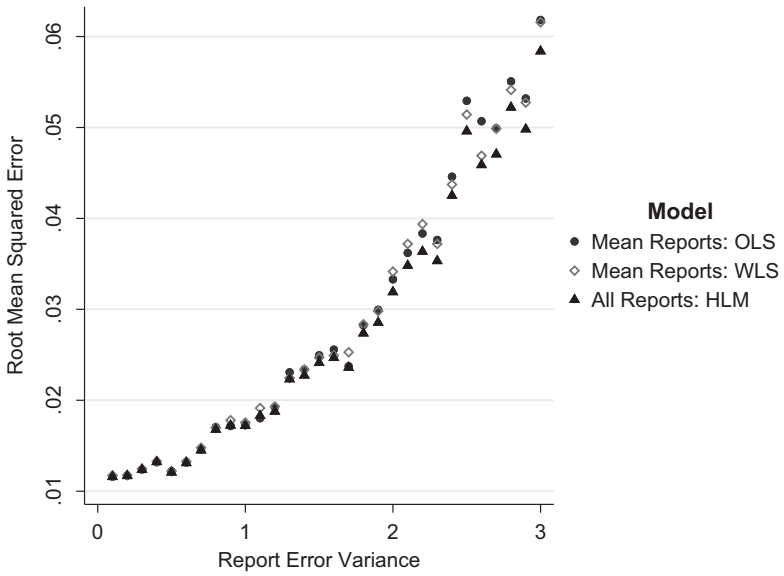


FIGURE 4 Mean Squared Error for $\hat{\beta}_1$ across Different Levels of Report-Level Variance



use information from all available reports. Specifically, we estimate two models that use the mean value of the reports $R_{r,s}$ for each event as the estimate of Y. In the estimation of the first model, *Mean Reports: OLS*, we regress this on X and estimate the model via ordinary least squares. Here, we are assuming that researchers have no information available to them besides the aggregated value of the outcome. If, however, researchers also know the number of reports for each event, they can estimate this model via WLS, relaxing the common variance assumption of OLS. Specifically, in our simulations, we weigh the observations by the number of reports used in generating the

higher-level outcomes—deflating the common variance by the number of reports per event.[14] This gives higher (lower) weight to those observations for which we have more (fewer) reports.

These results are reported in Figure 4. We see that each of the estimators performs similarly at low levels of report-level variance. As we increase the report-level variance—moving to the right across Figure 4—we see notable gains in the relative performance of HLM as compared to WLS and OLS. In short, the more unique information possessed within each report, the better HLM does, and the worse OLS and WLS do. This is because OLS and WLS neglect the information on variance leveraged in the HLM. Finally, across every condition evaluated in Figure 4, HLM weakly dominates OLS because HLM nests OLS. When there is no (or little) report-level variance, these estimation strategies will produce similar results. But as report variation increases, the gains realized from HLM become more pronounced.

In sum, compared to other multiple-report methods, HLM using report-level data offers a conservative approach for event outcomes: It does no worse and often better than traditional event-level analysis in recovering the effect of X on Y. This was found under data-generating conditions that are still relatively favorable to naïve strategies, as increased source-level error (or source bias) would further favor our HLMs. Moreover, as we demonstrate in the next section, HLM alone provides additional information on the data-generating process.

Illustration

To demonstrate the utility of these methods with empirical data, we evaluate these strategies using report-level data from a real project, the Mass Mobilization in Autocracies Database (MMAD).[15] Specifically, we analyze whether the scope of a protest (i.e., whether it addresses local issues or those of national importance) increases the number of participants. As discussed above, the MMAD provides details on each report generated about an event from the Associated Press, Agence France-Presse, and the BBC. As a result, we have multiple estimates of participation (i.e., turnout) in 42% of the events.

The measure of our dependent variable (*Turnout*) varies across the models: In *Single Report*, it is a single, randomly drawn report value; in *Max Report*, it is the maximum value of participation reported, and in *Mean Reports*, it is the average value of participation reported. Each of these models is estimated via least squares, allowing us to focus on the effect of different aggregation strategies. Next, we test whether supplying additional information about the variance in the accuracy of event estimates improves over these aggregated results. In WLS, the dependent variable is also the average value of participation reported; with observations now weighted by the number of reports used in generating each event-level mean. Finally, in HLM, we use the report values directly as the outcome, including random intercepts at both the event and country level. In all models, we include a single predictor, *Scope*, which is a categorical measure of the level the mass mobilization is directed at: 0 = *national*, 1 = *regional*, 2 = *local*.[16]

The results from these models are reported in Table 3. We consistently find that protests directed at local and regional issues have significantly lower levels of participation than those aimed at national issues (the reference category). However, there are two important things to note from the results. First, the use of different aggregation methods can produce substantially different results. For example, *Max Report* (Model 2) produces larger estimates than either *Single Report* or *Mean Reports*. Specifically, note the constant (i.e., when scope is national) in the *Max Report* model has a significantly higher estimate for participation than in all other models. In all, however, the results across these models are reasonably similar, indicating the the relationship between scope and participation is robust to different aggregation rules. Recall this type of model checking is *only* possible when researchers have report-level data and can compare different aggregation rules.

Second, the size of the effect is lower in the models that allow for nonconstant variance, WLS and HLM. While the gains from WLS are minimal, the HLM coefficients (Models 5 and 6) are substantively lower than the other models considered—that is, the gains are not just due to efficiency. More importantly, we obtain additional information about the determinants of participation. In particular, note that the variance across events is over 2, and the variance across reports of the same event is over 0.5 (both on the log scale). Of the 2,415 events for which we had multiple reports, the median difference between the maximum and minimum estimates was 300 participants—information that is neglected in the OLS estimates (Models 1–3), but accounted for in the HLM

[14] The results reported for WLS use the residuals squared as the weighting scheme, which performed better than using analytic weights.

[15] We cannot undertake an analog to the *Best Report* model(s) from the simulations, as we do not know the true values of the variables.

[16] In instances of conflicting reports on the event scope, we use the broadest reported scope (the minimum value) as the event value.

TABLE 3 Model(s) of Protest Participation (Logged) Using MMAD, 2003–12

	OLS Single Report (1)	OLS Max Report (2)	OLS Mean Reports (3)	WLS Mean Reports (4)	HLM All Reports (5)	HLM All Reports (6)
Scope (Regional)	−0.403	−0.488	−0.400	−0.367	−0.299	−0.277
	(0.080)	(0.083)	(0.078)	(0.076)	(0.070)	(0.070)
Scope (Local)	−0.402	−0.512	−0.404	−0.372	−0.378	−0.345
	(0.078)	(0.081)	(0.075)	(0.074)	(0.075)	(0.075)
Source (AFP)	—	—	—	—	—	−0.075
						(0.035)
Source (BBC)	—	—	—	—	—	−0.244
						(0.039)
Constant	6.378	6.554	6.364	6.306	6.284	6.409
	(0.023)	(0.024)	(0.023)	(0.023)	(0.106)	(0.109)
Event-Level Variance	—	—	—	—	2.126	2.107
					(0.048)	(0.048)
Country-Level Variance	—	—	—	—	0.574	0.559
					(0.120)	(0.112)
N	7,890	7,890	7,890	7,890	13,600	13,600

Note: The reference category for *Scope* is national and for *Source* is Associated Press. All findings above are statistically significant at conventional levels (i.e., p-value < .05).

results.[17] Furthermore, we can relax the restriction that sources only vary randomly by adding source dummies—AFP and BBC (with AP as the omitted category)—into the model (Model 6) to account for source-specific biases.[18] Not only do we observe a continued attenuation of the main effects (on *Scope*), but also that the reporting sources systematically differ from one another: AP estimates significantly higher turnout numbers than either the AFP or the BBC.[19] This may suggest relative market bias, with the AP either more likely to only report on high turnout events and/or to inflate turnout in a sensationalist manner. In all, the HLM results provide a much richer understanding of these data over what was possible in the OLS models.

This illustration shows the gains demonstrated in the simulations are also realized with empirical data. We want to highlight that a variety of additional methods can also be pursued with report-level data as outcomes. First, researchers could undertake alternative measurement modeling strategies (e.g., structural equation models, Bayesian measurement models), which are similar in their treatment of the event feature as latent (as in Fariss et al. 2017; Treier and Jackman 2008).[20] Bayesian measurement models, in particular, may be useful in contexts where researchers encounter mixed data types across reports—that is, one report indicates a specific value, whereas another gives a broad range. Second, while overimputation methods (such as those advocated in Blackwell, Honaker, and King 2017) can be applied to aggregated event data, these could be further improved with report-level data.

Here, we focus on the use of multilevel models (and specifically HLMs) to address measurement error in events data for three reasons. First, these methods are familiar to researchers—as they enjoy wide use in other contexts—and easy to implement. Second, although we focus on interval-valued outcomes, multilevel models are extremely flexible; therefore, our framework could be easily extended for the analysis of limited dependent variables. Finally, multilevel models allow researchers to easily incorporate additional structure to account for remaining heterogeneity (e.g., correlations in time and space) or biases (e.g., source-specific fixed effects). This could include source-level intercepts and/or covariates (e.g., distance

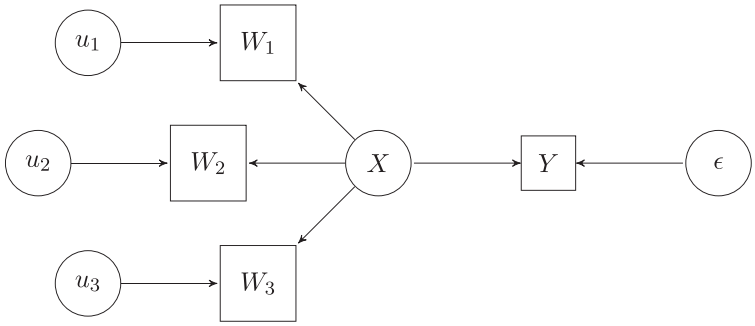
[17]The results presented in Model 5 are robust to the inclusion of source-level random effects.

[18]See Skrondal and Rabe-Hesketh (2004) for further discussion on the use of mixed effects models to account for source biases.

[19]AFP and BBC are also significantly different ($\chi^2 = 32.3$, pr > $\chi^2 = 0.000$).

[20]Though some of these strategies, such as item response theory (IRT), are less applicable here as the latent construct in event data *does* have a plantonic truth—that is, there is a true number of protesters—as opposed to unobservable latent constructs (e.g., political knowledge) where IRT is more appropriate (see Skrondal and Rabe-Hesketh 2004).

FIGURE 5 Report-Level Data and Predictor Error



Note: Researchers do not directly observe a predictor of interest X, but several reports W of the predictor with error u. Square nodes are observed, and circular nodes are unobserved.

from an event to a news bureau office) that produce systematic differences in measurement error. As such, they can provide insurance against both random and (known) systematic sources of measurement error in event reports.

Report-Level Data and Predictor Error

Data generated from media reports are widely used as predictors as well. For example, coup or coup success—coded from *New York Times* reports in Powell and Thyne (2011)—is included as a predictor in models of civil war duration, military spending, autocratic survival, economic sanctions, human rights violations, and pro-democracy protests. As above, researchers do not directly observe these events but instead possess one or more reports on them (see Figure 5). That is, researchers lack direct data on a predictor X but instead possess noisy reports W, which can be used as proxies. Treating these reports, or some transformation of these reports, as if they are the true predictor value will generally result in biased parameter estimates.

Measurement error in predictors, or error in variables, is elaborated in any introductory econometrics textbook, yet it proves useful for us to reintroduce some key concepts here. Consider a simple linear regression in which the predictor X suffers from classical measurement error:

$$W = X + u, \quad \text{where} \quad u \sim \mathcal{N}\left(0, \sigma_u^2\right), \tag{4}$$

where W is a noisy measure of the desired X. Regressing Y on W, instead of X, returns

$$\beta_W = \frac{\beta_X \sigma_X^2}{\sigma_X^2 + \sigma_u^2}, \tag{5}$$

with β_W underestimating the true relationship if σ_u^2 is not equal to zero, that is, if there is measurement error. This is classic attenuation bias.

Unless researchers can reliably assume that their predictors are error-free, accurate inference requires additional information. Occasionally, this can be achieved with the existing data if researchers are willing to make additional assumptions—see, for example, Blackwell, Honaker, and King's (2017) work on overimputation. However, where possible, researchers should seek supplementary data, which enable us to relax or test some of these assumptions (Manski 2011). Fortunately, with media-generated predictor data, this information is readily available in the reports. The key insight is again to recognize that we can treat the reports as individual noisy estimates (i.e., replicates) for the unobserved predictor. Treating reports as replicates opens up a variety of additional options for correcting measurement error: instrumental variable models, regression calibration, simulation extrapolation, and so on (Carroll et al. 2006). In the next section, we elaborate on one of these consistent estimation strategies.

Simulations

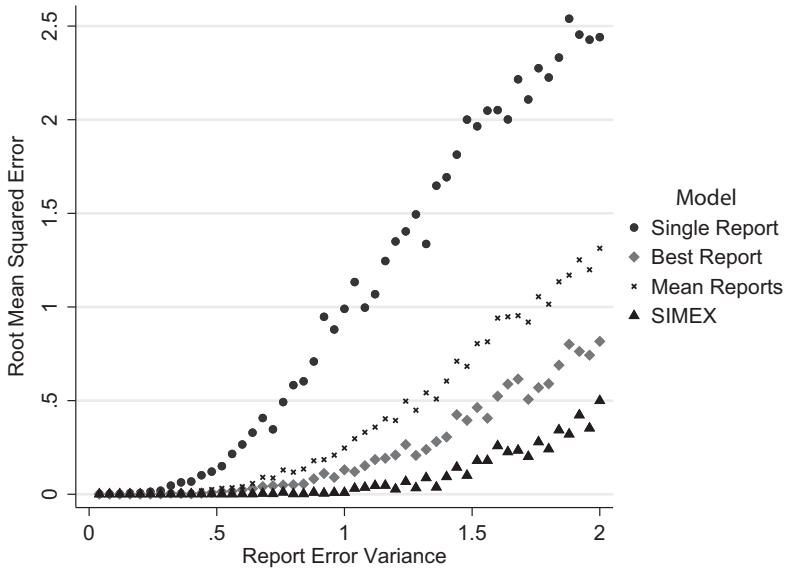
To demonstrate the consequences of using single reports or pre-aggregated event data, we undertake a series of experiments (similar to those in the previous section). Data on $N = 1{,}000$ outcomes are generated as

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{6}$$

where $X \sim \mathcal{N}(0, 1)$, $\epsilon \sim \mathcal{N}(0, 1)$, and $\beta_1 = 2$. However, X is only observed by the researcher via i noisy reports as

$$W_{r,s} = X + u_{r,s}, \quad \text{where} \quad u_{r,s} = e_r + \gamma_s, \tag{7}$$

FIGURE 6 Mean Squared Error for $\hat{\beta}_1$ across Different Levels of Report-Level Variance



with $e_r \sim \mathcal{N}(0, \sigma)$ and $\gamma_s \sim \mathcal{N}(0, \tau)$. The term σ determines the extent of the measurement error in the report, with $\sigma = 0$ indicating no error. In the simulations, we vary σ between 0.1 (almost no measurement error) to 2.0, holding all other conditions fixed. As before, we also randomly vary the number of reports generated for each respective event, allowing for a minimum of 0 and a maximum of 3.[21]

For this analysis, we confine attention to four empirical strategies. Three of these parallel the aggregation methods considered in the previous section: *Single Source*, *Best Report*, and *Mean Reports*. These are exactly as previously discussed, except now we include them as right-hand-side predictors in place of X. As before, *Single Report* should do poorly, *Mean Reports* should improve over this (via the law of large numbers), and *Best Report* should do better when sufficiently accurate. However, none of these strategies fully utilizes all of the available information provided at the report level: The variance in report accuracy means that event-level proxies (e.g., a mean) are more reliable for some observations than others. Therefore, our final strategy, simulation extrapolation (SIMEX), attempts to leverage this information to produce even better estimates.

SIMEX is an approximately consistent measurement error bias-correction method (Carroll et al. 2006; Wang et al. 1998). It adds additional measurement error to the replicate mean via simulation, fits a line to parameter estimates under these conditions, and then extrapolates back to the case of no measurement error. To see how this works, consider how Equation (5) changes when u has a larger variance: The second term in the denominator becomes larger, and β_W tends closer to zero. More generally, we could write this as

$$E(\beta_W|\zeta) = \frac{\beta_X \sigma_X^2}{\sigma_X^2 + (1+\zeta)\sigma_u^2}, \tag{8}$$

where $\zeta > 0$ and ζ represents the increase in the error variance. Using increasing values for ζ, we can simulate increasingly noisy proxies for X. Next, we estimate our model, simply substituting these synthetic values in for X, and store the coefficients. This allows us to estimate the rate of degradation in the coefficient estimates from additional error and ultimately infer back to the case of no measurement error ($\zeta = -1$). The intuition for this may be better conveyed geometrically, as we will do in the applied illustration below.

The results for the simulated experiments are given in Figure 6, which demonstrates that SIMEX weakly dominates the alternative estimators. As expected, the more the reports vary in their estimates (moving rightward in Figure 6), the worse *Single Report* performs. This reflects that any single estimate is no longer representative of the true value, with some higher and some lower. The performance of *Mean Reports* does not degrade as quickly, but it

[21]The report selection procedure is given in Equation (3), with W in place of D.

also does poorly as report variance grows unless one has a large number of reports for each event. Even *Best Report* with perfect report selection (100%) does poorly under high levels of report error variance.

That even the *Best Report* strategy performs poorly helps to demonstrate that errors in the reports are the fundamental issue, not errors made by data producers. That is, given sufficient report-level variance, even the most well-informed data producers cannot produce a single best data set that offsets measurement error. What we can do, however, is use the reports as replicates in a strategy such as SIMEX. In Figure 6, we see that SIMEX performs well under all the conditions evaluated, weakly dominating the other strategies. Specifically, the gains to SIMEX become sizable for report error variance greater than 0.2, which is a correlation of 0.98 between the X and the report. This means gains can be realized with even minimal reporting error.

Illustration

To illustrate this with empirical data, we again utilize the MMAD data set on protest turnout, using it as a predictor in a model of the level of violence employed. *Violence* is a binary measure for which 0 indicates no reports of physical violence and 1 indicates violence (e.g., people injured, people killed).[22] We focus on two main estimation strategies. First, two naïve logistic regression models, one with a single report of participation (*Single Report*) and the other using the average across reports (*Mean Reports*), Models 1 and 2 respectively. Second, logistic models with the SIMEX alogrithm used to correct for measurement error (Model 3). In the *SIMEX* model, the report values are used as replicate measures of the unobserved true participation. Standard errors are then calculated via nonparametric bootstrap.

The results are presented in Table 4. Our predictor, the level of participation, is positively and significantly related to the use of violence in a protest in all models. In addition, we see the expected attenuation effect when measurement error is not accounted for. In Model 1, the *Single Report* strategy with the most measurement error in the simulations produces the smallest estimate. When we use more information in *Mean Reports* (Model 2), we see an increase in the coefficient estimate as expected. However, the largest gains are realized in Model 3, where we use SIMEX to account for possible measurement error. In sum, our results show researchers would be consistently underestimating the effect of participation by using aggregated predictor data.[23] With report-level data, researchers can instead exploit the SIMEX algorithm and correct for this attenuation.[24]

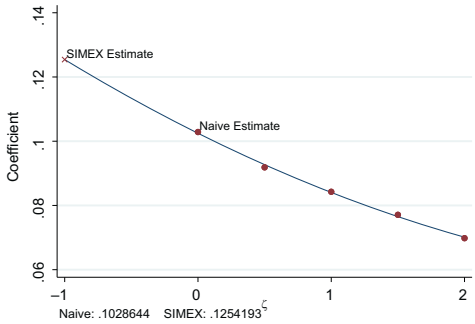
These gains can also be visualized as in Figure 7, which plots the results from Models 2 and 3. This also helps demonstrate the SIMEX procedure geometrically. First, the initial naïve estimate is obtained; then additional error is added iteratively to the predictor, with the effect

TABLE 4 Model(s) of the Use of Violence during Protest Using MMAD, 2003-12

	Logit Single Report (1)	Logit Mean Reports (2)	Logit SIMEX (3)
Participation (ln)	0.099	0.103	0.125
	(0.025)	(0.026)	(0.025)
Constant	−3.520	−3.525	−3.670
	(0.174)	(0.177)	(0.171)
N	7,865	7,887	7,887

Note: All findings above are statistically significant at conventional levels (i.e., p-value $< .05$).

FIGURE 7 Coefficient Estimates from Model 2 (Mean) and Model 3 (SIMEX)



[22] In Online Appendix C we show our results do not change if we only include explicit reports of no violence—rather than no mention of violence—as zeros. This is another benefit of possessing report-level data, as we can better guard against the "false zeroes" problem by estimating models that include reported zeroes, or include all non-reports as zeroes, and evaluate whether or findings are robust to both.

[23] We also calculate the marginal effect of participation on violence, consistent with the coefficient results we find that the naïve aggregation methods produce significantly lower values than the SIMEX model—respectively, 0.0049, 0.0052, and 0.0063.

[24] In SI Appendix C, we show how researchers could instead generate source-level (e.g., AP, AFP, BBC) means and then use these as predictor values of inputs for the SIMEX. The results (Appendix C, Table 3) are consistent with our presentation here.

estimate being calculated under each condition. From these results, we can fit a line (here, quadratically) and extrapolate back to the condition of no measurement error (the final SIMEX estimate). In sum, given replicate measures (i.e., multiple reports), we can add additional simulated error to draw inferences on the relationship between an error-free predictor and the outcome.

Conclusion

Media-generated event data are widely and increasingly used across many issue areas within political science. Such data are frequently the best available alternative for research into politically sensitive or divisive topics, for which gold standard data are difficult or impossible to acquire. While often necessary, it is important to be clear about the limitations of these data and to seek out ways to improve analysis utilizing them.

In the preceding, we have focused on an often overlooked step in the generation of event data: report aggregation. All event data sets currently used in applied research necessarily aggregate information from the report level to the event level. This pre-aggregation is not costless, as it can affect coding transparency and the robustness of results. Therefore, we argue that data should be provided to researchers at the report level. First, this would allow the researcher to empirically test the consequences of different aggregation rules, ensuring results are robust to different choices. Second, the report itself is the quantity we are able to collect data on. As such, our empirical models should parallel this aspect of the data-generating process. That is, we have reports from which we draw inferences on features of the event, and then we—often implicitly—use those estimates to draw inferences on causes of the event. Drawing accurate inferences on event causes requires introducing our uncertainty over the report estimates explicitly into our models.

Here, we have shown that statistical gains can be achieved by using report-level data directly in estimation. We hope that this will serve to advance a conversation on improving strategies for accounting for measurement error in event data, not end one. In the supporting information, we offer a general set of best practices for report data collection and analysis. We also elaborate on two areas that demand greater attention in future work. First, we consider adapting methods for analyses that use event data as both outcome and predictor. This is more complicated, as (1) methods combining error on both sides remain underdeveloped, and (2) it risks common-source bias—that is, differential measurement error induced by using a common source of information. Second, we explore improving our understanding of the *report-generating process* as distinct from the *event-generating process*. As we have discussed here, the data we observe are jointly determined by these two processes. Thus, our ability to understand *event causes* requires both an improved understanding of *report causes* and systematic data that enable us to incorporate these understandings into our models. This will enable improved analysis in any issue domain using media-generated data.

References

Aronson, Jay, Baruch Fischhoff, and Taylor Seybolt. 2013. "Moving Toward More Accurate Civilian Casuality Counts." In *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*, ed. Taylor Seybolt, Jay Aronson, and Baruch Fischhoff. Oxford: Oxford University Press 285–298.

Balcells, Laia, and Christopher M. Sullivan. 2018. "New Findings from Conflict Archives: An Introduction and Methodological Framework." *Journal of Peace Research* 55(2): 137–46.

Ball, Patrick, Jana Asher, David Sulmont, and Daniel Manrique. 2003. "How Many Peruvians Have Died? An Estimate of the Total Number of Victims Killed or Disappeared in the Armed Internal Conflict between 1980 and 2000." Report from the American Association for the Advancement of Science., Washington D.C. URL: https://www.hrdag.org/wp-content/uploads/2013/02/aaas_peru_5.pdf.

Baum, Matthew A., and Yuri M. Zhukov. 2015. "Filtering Revolution: Reporting Bias in International Newspaper Coverage of the Libyan Civil War." *Journal of Peace Research* 52(3): 384–400.

Blackwell, Matthew, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods and Research* 46(3): 303–41.

Buonaccorsi, John P. 1996. "Measurement Error in the Response in the General Linear Model." *Journal of the American Statistical Association* 91(434): 633–42.

Buonaccorsi, John P. 2010. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: CRC Press.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL: CRC Press.

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): 223–40.

Earl, Jennifer, Andrew Martin, John D. McCarthy, and Sarah A. Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30: 65–80.

Fariss, Christopher J., Charles D. Crabtree, Therese Anders, Zachary M. Jones, Fridolin J. Linder, and Jonathan N.

Markowitz. 2017. "Latent Estimation of GDP, GDP per Capita, and Population from Historic and Contemporary Sources." https://arxiv.org/pdf/1706.01099.pdf.

Hegre, Håvard, Gudrun Østby, and Clionadh Raleigh. 2009. "Poverty and Civil War Events: A Disaggregated Study of Liberia." *Journal of Peace Research* 53(4): 598–623.

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.

Hendrix, Cullen S., and Idean Salehyan. 2017. "A House Divided: Threat Perception, Military Factionalism, and Repression in Africa." *Journal of Conflict Resolution* 61(8): 1653–81.

Manski, Charles F. 2011. "Policy Analysis with Incredible Certitude." *The Economic Journal* 121(554): F261–89.

Mosley, Layna. 2011. "Replication Data for: Collective Labor Rights Dataset." https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/15590.

Powell, Jonathan M., and Clayton L. Thyne. 2011. "Global Instances of Coups from 1950 to 2010: A New Dataset." *Journal of Peace Research* 48(2): 249–259.

Salehyan, Idean. 2015. "Best Practices in the Collection of Conflict Data." *Journal of Peace Research* 52(1): 105–9.

Schrodt, Philip A. 2012. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38(4): 546–69.

Schrodt, Philip A., and Deborah J. Gerner. 1994. "Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982–92." *American Journal of Political Science* 38(3): 825–54.

Schutte, Sebastian, and Nils B. Weidmann. 2011. "Diffusion Patterns of Violence in Civil Wars." *Political Geography* 30(3): 143–52.

Seybolt, Taylor B., Jay D. Aronson, and Baruch Fischhoff. 2013. *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict.* Oxford: Oxford University Press.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: CRC Press.

Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46(1): 218–37.

Sullivan, Christopher M. 2016. "Undermining Resistance: Mobilization, Repression, and the Enforcement of Political Order." *Journal of Conflict Resolution* 60(7): 1163–90.

Treier, Shawn, and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1): 201–17.

Wang, Naisyin, Xihong Lin, Roberto G. Gutierrez, and Raymond J. Carroll. 1998. "Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models." *Journal of the American Statistical Association* 93(441): 249–61.

Weidmann, Nils B. 2015. "On the Accuracy of Media-Based Conflict Event Data." *Journal of Conflict Resolution* 59(6): 1129–49.

Weidmann, Nils B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60(1): 206–18.

Weidmann, Nils B., and Espen Geelmuyden Rød. 2015. "Making Uncertainty Explicit: Separating Reports and Events in the Coding of Violence and Contention." *Journal of Peace Research* 52(1): 125–28.

Weidmann, Nils B., and Espen Geelmuyden Rød. The Internet and Political Protest in Autocracies. Chapter 4. Oxford University Press, forthcoming.