How events enter (or not) data sets: the pitfalls and guidelines of using newspapers in the study of conflict
Demarest, L.; Langer, A.

Citation

How Events Enter (or Not) Data Sets: The Pitfalls and Guidelines of Using Newspapers in the Study of Conflict

Leila Demarest[1] and Arnim Langer[2]

Abstract

While conflict event data sets are increasingly used in contemporary conflict research, important concerns persist regarding the quality of the collected data. Such concerns are not necessarily new. Yet, because the methodological debate and evidence on potential errors remains scattered across different subdisciplines of social sciences, there is little consensus concerning proper reporting practices in codebooks, how best to deal with the different types of errors, and which types of errors should be prioritised. In this article, we introduce a new analytical framework—that is, the Total Event Error (TEE) framework—which aims to elucidate the methodological challenges and errors that may affect whether and how events are entered into conflict event data sets, drawing on different fields of study. Potential errors are diverse and may range from errors arising from the rationale of the media source (e.g., selection of certain types of events into the news) to errors occurring during the data collection process or the analysis phase. Based on

[1] Department of Political Science, Leiden University, the Netherlands
[2] Centre for Research on Peace and Development (CRPD), University of Leuven, Belgium

Corresponding Author:
Leila Demarest, Institute of Political Science, Leiden University, Pieter de la Court-building, Wassenaarseweg 52, 2333 AK Leiden, the Netherlands.
Email: l.demarest@fsw.leidenuniv.nl

the TEE framework, we propose a set of strategies to mitigate errors associated with the construction and use of conflict event data sets. We also identify a number of important avenues for future research concerning the methodology of creating conflict event data sets.

Introduction

With the quantitative turn in peace and conflict studies (e.g., Collier and Hoeffler 2002; Fearon and Laitin 2003; Hirshleifer 1994:3), the collection of detailed data on conflict events, actors, and casualty numbers has spurred many research projects in the field. The development of conflict event data sets has also accelerated in recent years. Major trends in new data projects include an increased focus on disaggregated conflict events—both in time and in space—as well as a focus on low-level forms of conflict such as protests, as opposed to civil war (Bernauer and Gleditsch 2012:375-78; Gleditsch et al. 2014:303-5, 308-9). News reports have been the most important source of data on conflict events, as they are widely available and often accessible at low cost. However, the widespread use of media reports as an empirical source raises important concerns about the quality of event data.

The collection of political event data has a long history, both in the social movement (e.g., Eisinger 1973) and in the international relations literature (e.g., Azar 1980; McClelland 1976). Concerns about the validity and reliability of media data for political science research are hence not necessarily new (e.g., Danzger 1975; Franzosi 1987). Nonetheless, the increased availability of (online) media sources and the development of new data sets has spurred new debates in this field. An important characteristic of many new data sets is their geographical focus on the developing world. Examples include the Armed Conflict Location and Event Dataset (ACLED; Raleigh et al. 2010), the Social Conflict in Africa Database (SCAD; Salehyan et al. 2012), the Urban Social Disturbance in Africa and Asia (USDAA) data set (Urdal 2008), the UCDP Georeferenced Event Dataset (UCDP GED; Sundberg and Melander 2013),[1] the Global Terrorism Database (LaFree and Dugan 2007), Political Instability Task Force (PITF) Worldwide Atrocities Dataset (Schrodt and Ulfelder 2016), the Mass Mobilization in Autocracies Database (Weidmann and Rød 2015), and the Konstanz One-Sided Violence Event Dataset (Schneider and Bussmann 2013). By contrast, much of the

methodological debate and evidence with respect to the use of media reports to construct event data is found in Western-focused social movement research (e.g., Earl et al. 2004; Hutter 2014) as well as in communications studies (Galtung and Ruge 1965; Harcup and O'Neill 2016; Krippendorff 2013). Further, while in recent years, different methodological challenges associated with generating new conflict event data sets have been critically assessed (e.g., Eck 2012; Salehyan 2015; Weidmann 2015, 2016), this field of study stands to benefit from further systematization of research findings.

To this aim, the current article introduces a new analytical framework that captures the methodological challenges of using news reports for generating conflict event data and recognizes a broad range of errors that may affect whether and how events are entered into data sets. The Total Event Error (TEE) framework draws on insights from the survey research literature and the Total Survey Error (TSE) framework. In analogy with Groves and colleagues (2004:41-63), we distinguish between measurement errors and errors of representation. The framework encompasses well-known forms of error mentioned in the literature, such as selection bias, which arises when newspapers deliberately select some events for publication, while leaving other events unreported (Earl et al. 2004:68-72; Jenkins and Maher 2016; Ortiz et al. 2005; Weidmann 2016). However, we also consider errors that are not necessarily caused by the rationale of media sources and have received much less attention in the literature. These errors arise during the data collection process, such as the coding of key variables, or in the analysis phase, when researchers make use of imputed values for missing data (e.g., the location of an event). Further, while bias, or a systematic difference between the measured value and the real value, is an important form of error, we also direct attention toward unreliability or random deviation from the real value, which undermines precision.

The TEE framework offers a bridge between methodological insights from conflict event studies and Western-focused social movement and communications studies. This has the advantage that insights, methods, and procedures that are common in these latter literatures are introduced and discussed with regard to conflict events in developing countries. We devote particular attention to the implications of focusing on developing contexts as opposed to Western contexts. Indeed, while Western-centred studies have commonly focused on protest events, we focus on a wider range of events, including protests, but also violent armed conflict events. Finally, we discuss and compare human as well as automated forms of data collection and coding. Although optimism has often been expressed with regard to the potential opportunities and advantages of automated coding (e.g., Bond et al. 1997;

King and Lowe 2003; Schrodt and van Brackle 2013), so far, it is not yet widely used in conflict studies. Illustratively, human coding is used by all data projects cited above. Arguably, the main reason for why human coding has remained the common practice is that in recent years, conflict scholars have aimed to construct conflict event data sets on the basis of increasingly complex information drawn from media reports (e.g., Hammond and Weidmann 2014). Having said this, automated coding has important advantages compared to human coding and is therefore likely to gain more relevance in conflict studies in the coming years.

As an analytical framework, TEE offers an important methodological basis for studies on conflict events and gives guidance to developers and users of old and new data sets. For developers, the TEE framework systematically sets out the different types of errors to be reflected upon in data codebooks, or articles introducing new data sets, and supports standardization of reporting practices in the field. Indeed, as the collection of event data and the use of media data have been taken up by different subdisciplines and areas of social science, the types of errors researchers are concerned with, or report on, appear to differ widely. As will also become clear from our discussion of the state of the art, relatively little is known concerning errors that may arise when collecting data on conflict events in the developing world. To fill this gap, new empirical research is necessary. On the basis of the TEE framework, we are able to identify a number of important avenues for future research. The TEE framework is also extremely useful for conflict event data users because it provides important insights concerning the range of errors one has to take into account when using a specific conflict event data set, and how these errors may potentially affect research findings.

In the following section, we develop the TEE framework and discuss in depth the measurement and representation errors that can arise during each step in the research process. Our discussion is supported by (necessarily eclectic) empirical examples drawn from literature. In the third section, based on the TEE framework, we introduce guidelines and strategies for event data collection and future research. The fourth section concludes.

The TEE Framework

The TEE framework is inspired by the well-known TSE framework used in survey research (Groves et al. 2004:41-63). In the TSE framework, measurement errors occur when the measured value deviates from the real value. This can arise from unclear question wording and answer scales, the presence of an interviewer, which inhibits the respondent from answering truthfully (i.e.,

social desirability bias), or the incorrect processing of data. Errors of representation occur when not all existing observations are sampled in the survey. An essential characteristic of a survey is the sampling of only a subsection of the population, implying that this form of error always occurs. What is important is that observations are sampled randomly. This randomness can be jeopardized by a flawed sampling frame, nonresponse, or data adjustments based on a flawed external source (e.g., an outdated census). Clearly, two forms of error can occur: bias, which causes a systematic deviation from the real value, and unreliability, which arises from random errors, making the results less precise.

The collection and use of event data resembles the survey process in important ways. An important similarity is that sampling is inherent to the process. By selecting media sources to capture conflict events, one is aware that not all events that have taken place are necessarily reported. The challenge arises from the nonrandom processes steering media event inclusion, a debate which can be related to concerns about nonresponse error in surveys. Like respondents, news sources and reports can present information in a biased way or they may simply not be able to provide the necessary information, leading to missing data. While the interviewer commonly plays a key role in the sampling of respondents for public opinion polls, the same is true for a coder in charge of sampling relevant events into a data set. Furthermore, unclear coding instructions or variable definitions and categories can lead to unreliable or biased data, as can unclear survey questions. For both types of data, researchers can attempt to validate data against an external source. This can be a census or medical records for surveys or police and nongovernmental organization (NGO) reports for event data. Lastly, in the analysis phase, researchers can choose to weight the data to compensate for nonresponse or biased selection or they can choose to impute missing data to preserve the number of cases in the analysis.

Figure 1 visualizes the TEE framework. Central to the figure are the research steps taken in event data collection and analysis. These steps are not necessarily sequentially taken and may interact in important ways. For example, the development of the codebook is not necessarily finalized before the coding process, as a coding pilot test helps in refining the codebook. In case of automated coding, the coder has no role or at least a much more limited one. The development of the codebook (or dictionary in automated applications) becomes all the more important. In addition, comparisons to nonmedia sources are not often realized, simply because of a lack of such external data. In line with Groves et al. (2004:48) study, we associate each research step with both measurement and representation error. Several
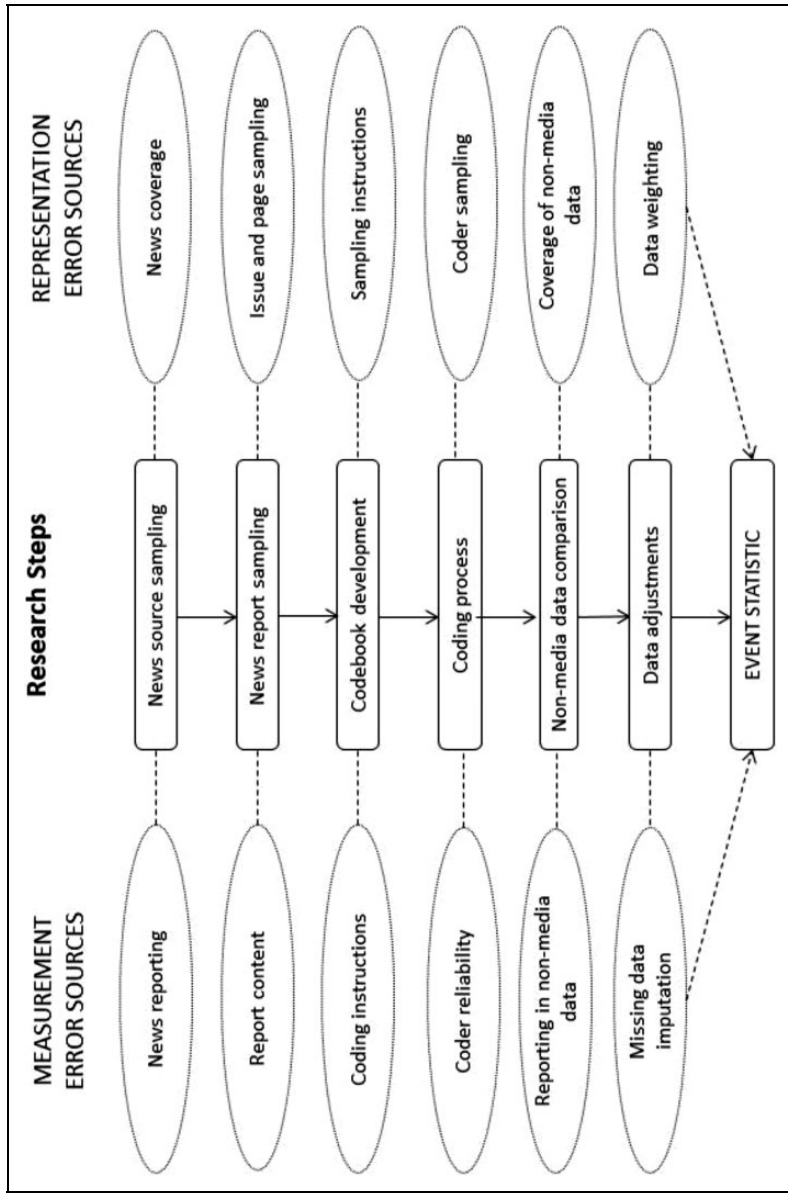
Figure 1. Total event error.

sources of error have been touched upon above, but they are addressed in greater depth in the following sections. We structure the discussion according to the research steps identified.

News Source Sampling

News coverage. News source sampling can give rise to both measurement and representation error. We start the discussion with representation error caused by news coverage effects, as this relates to the relatively well-known problem of selection bias (Earl et al. 2004:68-72; Jenkins and Maher 2016; Ortiz et al. 2005).[2] Nevertheless, while bias has been widely studied in the literature (Earl et al. 2004:68-72; Jenkins and Maher 2016; Ortiz et al. 2005), coverage effects can also be associated with unreliability, much as is the case with sampling error.

When deciding on collecting event data for specific types of conflict, time periods, and geographical settings, researchers first decide on the news source from which to extract data. This can be a newspaper, a news wire service, or even television and radio news. The choice of a news source implies that events included in the data set are dependent on media selection (or sampling) of events into the news. As a multitude of studies has shown, this selection is far from random. A dual problem is apparent: News source coverage can be determined by the characteristics of an event but also by the characteristics of the news source itself. The first is seen as a coverage effect common to different media outlets, but the second can be source-specific and underscores that the question from which news sources to extract data is an important one. We first discuss general, then source-specific selection effects.

The seminal paper by Galtung and Ruge (1965) on the presentation of the Congo, Cuba, and Cyprus crises in Norwegian newspapers set the basis for news value theory in communications science, which investigates the characteristics of an event that are likely to make it newsworthy (Harcup and O'Neill 2001). Galtung and Ruge propose 12 news factors that determine whether a foreign crisis event will be reported, including the event's amplitude or importance and the involvement of elite actors. Following their work, other communications scientists have investigated the news values that determine selection into the news media and have increased or reduced the number of relevant factors (e.g., Harcup and O'Neill 2001, 2016).

Social movement scholars focus specifically on protest events (Hutter 2014; Koopmans and Rucht 2002). Summarizing the findings of previous studies, Earl et al. (2004:69), Ortiz et al. (2005:398-400), Jenkins and Maher

(2016:45-46), and Hutter (2014:350-51) note that large-scale protest events with many participants, events characterized by violence (property or physical damage, police repression, arrests, etc.), events organized by movements with professional (public relations) staff, and events involving high-profile actors are all more likely to be reported. These findings are supported by comparisons of event inclusion between media sources but also by comparisons of media reports with external sources such as police records.

Representation error has been far less investigated with respect to conflict event data in the developing world. Recent interest in low-level conflicts, including protests in developing countries, can perhaps assume the same coverage preferences. For armed conflict events, we could assume that because of the level of violence, selection into the news is highly likely. However, the contexts in which armed conflicts arise are often different from the Western settings commonly investigated. Civil wars often erupt in rural areas away from the government's center of power (e.g., Kalyvas 2006:38-48), which has implications for the communications infrastructure present in the region. Furthermore, although armed conflict attracts journalistic attention, a climate of violence and infrastructure damage can obstruct event coverage. In this regard, a study by Weidmann (2016) is highly instructive. He compares a data set on armed conflict in Afghanistan collected by the U.S. military—and revealed by WikiLeaks—with the UCDP GED data set (which solely used media sources for this conflict) and finds that cell phone coverage significantly increases the likelihood of events being reported in the media, suggesting a systematic underrepresentation of events in remote rural areas.

Important source-specific selection effects are related to the ideological or political orientation of a news source, as well as its geographical scope (e.g., Davenport 2010:107-26). For example, in the analysis of protest events, several studies find that conservative newspapers underreport violent demonstrations to limit copycat behavior (for an overview see Ortiz et al. 2005:401). The second factor, the geographical scope of the news source, relates to whether a local, national, or international target audience is reached. We devote more attention to this issue here, as many recent data sets make use of multisource inventories, such as Factiva, LexisNexis, or Keesing's Record of World Events,[3] which rely to an important extent on international news wire services to code events occurring in a wide range of developing countries, including violent armed conflict as well as protests.

Local news sources can cover local conflicts more extensively than national sources, which implement an additional selection procedure. International sources have an even more stringent selection process. However, for

some events, such as ongoing armed conflict, professional international news wire services could potentially be more valuable than (disrupted) local media services. Exactly how selection bias plays out when the scales of conflict and news source scope interact is a highly relevant and perhaps insufficiently addressed empirical question. Several studies do indicate its importance. Herkenrath and Knoll (2011), for example, find that international newspapers report substantially less protest events than national newspapers in Argentina, Mexico, and Paraguay and that these differences are related to, among others things, the use of violence but also to a general difference in international media attention toward these three countries.

Bueno de Mesquita et al. (2015) developed a data set on political violence in Pakistan based on national newspapers and record a higher number of incidents than data sets relying on Factiva. Demarest and Langer (2018) compared data sets based on international versus national news sources on conflict events in Nigeria. They find that international sources underrepresent conflict events, in particular protest events. Both studies also find that relative underreporting affects the subnational distribution of events, an increasingly important research line in conflict studies (e.g., Gleditsch et al. 2014:303-05). Lastly, Barron and Sharpe (2008) used district-level news sources to capture violent events in Indonesia and show that these record more incidents than provincial newspapers and hence provide greater insights into local causes of conflict.

In general, multisource inventories are argued to be more reliable than single sources (Jenkins and Maher 2016:47-49), but it is important to keep in mind that multisource inventories do not include the "universe of media reports" (Ortiz et al. 2005:402). Especially for conflict in the developing world, it is important to consider that the international, English sources included in these inventories might not cover these settings sufficiently (Schrodt 2012:552-53). Automated coding procedures in principle are not sensitive to representation error, but they do require machine readable text and predominantly draw on international news wire reporting services such as Reuters and LexisNexis (e.g., Integrated Data for Events Analysis data set, Bond et al. 2003; Global Data on Events, Location and Tone [GDELT] data set, Leetaru and Schrodt 2013; Kansas Event Data System [KEDS] data set, Schrodt 2006), which is an important characteristic to consider. The use of local newspapers to investigate conflict events in developing countries is likely to emerge as an important research line in conflict studies, not in the least because local newspapers are increasingly available online (e.g., AllAfrica repository).[4] However, the use of local sources brings with it new challenges, for example, related to media ownership and state control

of media sources. So far, little systematic research appears to have been conducted to assess the impact of these issues on the quality of conflict event data.

News reporting. We now turn to the problem of measurement error arising from the news source, which concerns the information news sources report with regard to an event. This form of error can be linked to the concept of description bias (Earl et al. 2004:72-73). When discussing description bias problems, several scholars make a distinction between "hard news" and "soft news" and argue that the former is less subject to bias than the latter (Earl et al. 2004:72; Franzosi 1987:7; Raleigh et al. 2010:656).[5] Hard news is suggested to include the "who, what, when, where, and why of the event" (Earl et al. 2004:72), whereas soft news is said to include interpretations of causes and consequences, portrayals of the actors, and so on. We first discuss research on soft news dimensions and then focus on hard news. As argued below, the distinction made in the literature between hard and soft news is, however, not straightforward. Furthermore, not all reporting inaccuracies are necessarily signs of bias but can also indicate unreliability due to challenges for media sources to acquire certain types of information.

Soft news effects can be related to the concept of framing. Several definitions of framing exist; as an example, we cite Entman (1993:52): "To frame is to select some aspects of a perceived reality and make them more salient in a communicating context, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation." A substantial amount of research has focused on the way in which media represent protest actions, albeit predominantly focused on Western settings (e.g., Dardis 2006; McLeod and Hertog 1992). A major research line focuses on differences in framing according to the ideological profile (conservative or liberal) of the news source and whether conservative newspapers are more likely to depict protesters negatively (e.g., Chan and Lee 1984; Lee 2014; Weaver and Scacco 2012).

Although literature on the framing of conflict events provides interesting insights into the orientations of different news sources, it is less clear to what extent different representations of conflict can affect event data sets that focus on dates, locations, and actors. Nonetheless, the line between soft news and hard news is not necessarily clear-cut. A well-known example of a commonly contested "hard fact" is the number of participants at a protest, which can be exaggerated by activists or understated by police authorities (Day, Pinckney, and Chenoweth 2015:130). Furthermore, fatality estimates are often regarded as hard facts that are difficult to establish (Raleigh et al.

2010:656; Sundberg and Melander 2013:527). The source—police versus protesters or government versus rebels—that is preferred by a sampled newspaper can bias event data statistics.

Relatively few studies have compared external data with the reporting of hard facts in the media. McCarthy et al. (1999:117-26) compared data from police records with print (and electronic) media reports for protest events in Washington, DC. They found good correspondence for protest dates and purpose but weaker agreement concerning protest size. The latter could be related to bias or unreliability, however, as the analysis does not describe how the variables relate to each other. Weidmann (2015) investigated differences in the reporting of hard facts between the U.S. military data set on armed conflict events in Afghanistan and UCDP GED. He finds that for most events, the casualty numbers of the military data set fall within the low–high casualty estimate of the UCDP GED data set. There are more events for which UCDP GED gives a higher estimate though, which could indicate a slight bias toward reporting higher casualty numbers in news reports. There are also differences between UCDP GED and the military data set in the reported location of an event. Based on his analyses, Weidmann (2015:1143) argues that researchers should not use data for analyses below a range of 50 km. His research indicates that even a hard fact such as "location" is also not always reported reliably, in particular when considering armed conflict events. Although fatalities are considered difficult to establish reliably, Weidmann's research suggests that their reporting appears relatively free of error.

News coverage and reporting relate to some of the best-known errors described in the literature. Nonetheless, many studies focus on Western contexts and protest events and only to a lesser extent on developing contexts and events of violent (armed) conflict. While important principles and lessons can be drawn from social movement and communications studies, there is a clear need for more empirical research on these forms of errors in conflict studies. In the following sections, we turn to errors that are arguably less widely discussed in current scholarship. These errors do not necessarily arise from the workings of media sources but are more related to data collection procedures.

News Report Sampling

Issue and page sampling. While some researchers draw on all reports available from a specific source, others rely on the additional sampling of specific newspaper issues or pages (Earl et al. 2004:68; Krippendorff 2013:112-25). When it comes to recent conflict data sets (see Introduction), this additional sampling stage is not included, as they commonly rely on reports drawn from

multisource inventories, using key terms and date and country specifications. For studies relying on national or local newspapers, especially when a relatively extensive period is being studied, this additional sampling may be necessary to reduce coding costs. For example, for their seminal study on protest events in four Western-European countries, Kriesi et al. (1998:253-63) used one national newspaper per country but only the Monday edition. They covered the period from 1975 to 1989. Even if no systematic biases are associated with specific newspaper editions, this additional sampling engenders further unreliability. It is also possible to select only the first page of a newspaper issue, which could for instance reinforce bias toward the inclusion of high-profile events characterized by violence.

Report content. Journalistic or editorial preferences can also be an important source of error at the level of the news report. For example, Chojnacki et al. (2012:390-92) draw attention to the fact that news reports often quote sources that have incentives to provide biased information. They use the example of a report in which a rebel leader claimed to have killed 30 government soldiers. In their data set (Event Data on Conflict and Security [EDACS]), they created an additional variable, indicating that the information might be biased if doubtful sources are used.

In addition to the biases that can arise from reporting preferences, news reports themselves can be important sources of unreliability. First, reports on events can be detailed or vague. Some reports might provide information on the size of a group of protesters, whereas another report on the same event might only mention that the protest occurred. Similarly, the capture of territory by a rebel group can be reported but not necessarily whether there were any casualties. In some cases, multiple reports can provide valuable additional information, yet for others, vague reports might be the sole source of information and a substantial degree of missing data can result. The newsworthiness of an event can also affect the depth of reporting and the length of the news article devoted to it. For example, a large-scale protest can attract more news attention than an event of limited size, and hence, more information on the event might also be reported. Nonetheless, while some events can gain strong news attention, such as grave human rights abuses in armed conflict, the "fog of war" can also prevent the collection of reliable information.

Second, reports can also explicitly cast doubts on whether and how an event occurred, on the identity of the actors, or on the validity of a casualty estimate. Reports can, for example, state that the identity of attackers or

suspected rebels is uncertain. These forms of measurement error can only be captured if such indicator variables are included in the codebook.

Third, coding challenges can arise from conflicting reports. While the incompleteness of news reports leads many researchers to draw from multiple reports to construct event data variables, this can also raise additional questions concerning the way in which reports are combined (Weidmann and Rød 2015:125-26). A crucial problem arises when information is inconsistent. Some data sets provide instructions to coders to aggregate the information in particular ways. For example, SCAD states that in the case of multiple casualty estimates, the mean is taken (Codebook version 3.1.), whereas ACLED states that the lowest number should be used (Raleigh and Dowd 2017:20). Other solutions to conflicting reports suggest coding each report individually. Based on their work on protest events for the Nonviolent and Violent Campaigns and Outcomes data set, Day et al. (2015:130-31) recommend the coding of different reports, together with including a metric ambiguity range variable in the final event data set.

Similarly, Weidmann and Rød (2015) propose the creation of an intermediate data set, which includes the event coding by news report, and an event data set, which aggregates the information across reports. As all reporting information is provided, aggregation rules (mean, minimum, etc.) can be altered. Coding news reports separately can increase transparency and replicability, as opposed to allowing coders to aggregate news reports themselves. This coding choice can also have important implications for the monitoring of the coding process and intercoder reliability scores (see below). Coding news reports separately can, however, increase research costs.

Codebook Development

Sampling instructions. Codebook instructions are crucial to avoid coder confusion and to support consistent sampling as well as coding of relevant events. When using machine coding, the dictionary and coding program determine selection and coding of cases into the data set based on the identification of relevant actors, wordings, and so on, rather than a coder.[6] Generally, codebooks and dictionaries are revised after an initial coding test phase, in which potential sources of error are revealed. Sampling instructions are an important concern: Which events should be included in the data set and which should be excluded?

When developing instructions for human coders, researchers can either adopt a definition or provide a list of eligible events (e.g., Kriesi et al. 1998:263-69). Many conflict event data sets mainly rely on event definitions,

but a potential caveat is that the stricter the definition, the more difficult it becomes to consistently code vague reports of events. Reports do not always give details on actors, which actor used violence, or the number of participants at an event, for example, which can create confusion and sampling inconsistencies when categorization requires this information. It can also be important to include instructions on how to handle cases for which a report casts doubts on its occurrence or eligibility for inclusion.

When sampling events from online repositories, the same concerns apply. In databases such as LexisNexis, one can develop a search string of relevant key words and apply these to extract news stories about a specific topic or event. Afterward, a subsample can be manually verified by coders to select the usability and efficiency of the search string and the amount of "noise." Nevertheless, a coder's decision to include or exclude events still requires consistency and replicability and consideration of the aforementioned issues. A news report that includes relevant key words such as "violence" might report more than one event, for example, all of which need to be sampled consistently. Furthermore, the use of search strings does not assure that all events sampled from a news source are also sampled by using specific terms. Although search strings often include many key terms, some events can still be overlooked.

The issues of noise and the overlooking of events are also a major concern when using automated coding procedures. It is useful, however, to first point out the benefits of machine coding. While the development of dictionaries is time-consuming, including as many key verbs and phrases, variations, names of actors (e.g., United States, US, USA, President Trump) as possible, once developed, they offer the potential to go through large volumes of data in seconds (Bond et al. 1997; Schrodt and Van Brackle 2013). Further, a revision of the dictionary does not result in a time-consuming recoding process. Instead, the program can just rerun on the same data with the revised dictionary. Finally, dictionaries can be shared between researchers and be used for new projects. A major point of discussion is however whether machine coding is able to identify the "right" events and whether these events are coded correctly (see below), with human coding often taken as the standard.

The ability of machine coding procedures to include a sufficient high number of relevant events ("recall"), while at the same excluding irrelevant events ("precision")—events related to sports competitions are common false positives—is an important sampling challenge.[7] Several researchers have empirically investigated recall and/or precision for machine coding applications compared to a training set developed by human coders. For instance, Bond et al. (1997) find that the original KEDS's sparse parsing

program[8] performs as least as well as (new) human coders in identifying relevant events (around 80 percent). King and Lowe (2003) test the VRA reader and find that it performs as well as human coders for recall (93 percent correct) but less for precision (23 percent correct). Overall, they are positive about the potential of machine coding, however.

Besides comparisons with human coders, there have also been comparisons between programs which are continually developing. Boschee, Natarajan, and Weischedel (2013) compare the TABARI program developed by Schrodt as a follow-up to the original KEDS program and find that with regard to recall and precision, its sparse parsing procedure is significantly outperformed by the BBN SERIF program that relies on natural language processing.[9] Most recently, Croicu and Weidmann (2015) developed a machine learning classifier system that shows recall and precision percentages of around 90 and 50, respectively, again as compared to human coders. Heap et al. (2017) propose a joint human/machine process for the selection of relevant text by supervised machine learning to improve recall and precision. Besides natural language processing and machine learning, another area of progress in automated coding lies with conditional random fields (Schrodt and Van Brackle 2013:38; Stepinski, Stoll, and Subramanian 2006).

It appears that automated coding has important and increasing benefits for event sampling. There continue to be a number of challenges to consider, however. The first crucial challenge concerns duplication or the inclusion of the same event into the data set multiple times (Bond et al. 2003:737-38; Schrodt and Van Brackle 2013). There is no real automatic procedure yet to filter out duplicates, except to discard events with the same time, location, actors, and so on. Human review of the data set can be required to exclude further duplicates and can still be a costly exercise when considering large volumes of data. Another challenge concerns language, as most dictionaries and applications predominately focus on the English language (Schrodt and Van Brackle 2013:45), while extensions to other languages can lead to the inclusion of more diverse and non-Western sources. Nonetheless, the use of English is also not uniform, and specific word choices and sentence structures can also vary across regions or countries and can be more pronounced for domestic than international events (Schrodt, Simpson, and Gerner 2001). Even the news source itself can vary in language use (Boschee et al. 2013).

Automated coding has primarily been used for the collection of political event data in the field of international relations (e.g., Schrodt 2006). Increasingly, the use of automated coding is also used to investigate domestic conflicts, including in developing contexts (e.g., Leetaru and Schrodt 2013).

This implies that the challenges with regard to dictionary construction described above are becoming increasingly pertinent to deal with, both when concerning the selection of events and the coding of events, as will be discussed below.

Coding instructions. Unclear coding instructions can create representation errors as well as measurement errors. Again, the problem of defining events arises. For example, the USDAA codebook includes 12 event definitions, but it is argued that these conflict types "are by no means mutually exclusive categories. [. . .] While we have tried to be consistent in the coding of such events, one should be careful in treating the categories as clearly distinguishable phenomena" (Urdal 2008:11). This problem stems from missing or conflicting information in event reports. In some data sets, for example, the mentioning of an association behind the protest can make the difference between categorization as a spontaneous or as an organized protest (e.g., SCAD).[10] Yet, this can also be influenced by the depth of reporting.

When developing the codebook, researchers potentially have to choose between very generic categories of events, actors, and so on, which can be coded reliably, or very specific categories, for which coding is more unreliable. This is an important trade-off to be made. While broad or generic categories might create more consistency, they might not provide the level of information precision that researchers strive for. A generic actor category such as "attackers" might be coded very reliably, for example, but one would also want to know, where possible, whether the attackers were particular rebel groups or ethnic militias, political parties, and so on. Unfortunately, the need for detailed event information to pursue particular research questions is not always accommodated by the information provided in media sources.

For automated coding, the complexity of event coding is not only challenged by the information available in news reports but also by the dictionary and the nuances predefined sentence structures can capture. Bond et al. (1997) also analyzed event categorization besides event sampling and find again that machine coding performs similar to human coding. King and Lowe (2003) have similar findings but also show that more general event classifications are coded more reliably than detailed ones. The fact that detailed event definitions are not always workable is also discussed by Schrodt and Van Brackle (2013:33).

In general, automated coding is deemed to work better when the variables that need to be extracted are not too complex. One challenge here is that the field of peace and conflict studies is increasingly moving toward more

complex event definitions and characteristics, as well as detailed collection of time and location information. As discussed above, subnational location information for events is increasingly sought after in empirical research, yet automated coding is argued to work better on the country level (Bond et al. 2003:739; Schrodt and Van Brackle 2013:46). Hammond and Weidmann (2014), for instance, argue that the GDELT data set should be used with caution for subnational analyses as it differs substantially from human coding and seems to show a bias toward country capitals. Hickler and Wiesel (2012) are more optimistic when comparing spatial information for human and machine coded data in the framework of the EDACS data set, yet concerns are still raised.

When human coding is used, the development of the codebook is an important start, yet how it is implemented is to a large extent the responsibility of the coders. Machine coding rules out coders or gives them a more limited (supervising) role (e.g., Heap et al. 2017). In the following section, we will focus on errors arising from the coder in a typical human coding project. Interestingly, even though machine coding is commonly compared to a human coding benchmark, human coding itself is also subjected to substantial errors. This is indeed the core argument of Bond et al. (1997:555) who early on lamented the poor quality of human coding.

Coding Process

Coder sampling. Following codebook instructions, coders sample events into the data set and extract information on key variables. Thus, the coder can also be a source of representation error and measurement error, and both unreliability and bias can arise. When sampling, coders can overlook events completely at random due to, for example, inattentiveness. Bias occurs when coders routinely overlook certain events or regularly misinterpret instructions on what constitutes a relevant event. Unfortunately, it is likely that smaller, low-scale events more often go unnoticed than high-profile events announced in headlines (e.g., Kriesi et al. 1998:270), which is why coder sampling error can potentially reinforce selection bias. Coder sampling error is not often measured (or reported), but some researchers have attempted to quantify it. In their work on social movements in four Western-European countries, Kriesi et al. (1998:270) report that in paired comparisons, around 60 percent of protest events were registered by both coders. A follow-up project reached about 70 percent identification agreement between coders (Hutter 2014:355). Although they used a different data source than news reports—reports from the United Nations Secretary General on peacekeeping

operations—Ruggeri, Gizelis, and Dorussen (2011:348-51) also note severe coder sampling error. They find that independent coders only double-identified 18–41 percent of relevant events. This necessitated the research team switching strategies and having the team leaders identify and highlight relevant events, which were then coded by the assistants.

Coder reliability. For research that makes use of media content analysis, the calculation of intercoder reliability to indicate measurement error is regarded as a methodological imperative in communications science (Krippendorff 2013:272-73).[11] This imperative has also made its way into protest event analyses in (Western) social movement studies (Hutter 2014:354-55). However, many conflict event data sets focusing on the developing world do not report such measurements (Ruggeri et al. 2011:356-59; Salehyan 2015:107-08). By conducting intercoder reliability tests, however, one can check whether the same measurement instrument (the codebook) leads independent coders to reach similar results (Krippendorff 2013:273-75). Common measures are Krippendorff's α and Cohen's κ, which both correct for chance agreement by weighing inconsistency in less frequent response categories more heavily in the final coefficient. Intercoder reliability checks can be used to refine the codebook or select the "better" coders after a pilot stage. It is recommended to conduct tests regularly throughout the coding process as only conducting postdata collection tests can reveal the need to discard or recode a substantial amount of data. The tests can be conducted on a small subset of the data (5–10 percent).

When interpreting intercoder reliability, it is also important to be aware that low intercoder reliability can arise if each coder makes random errors (coder unreliability) or if each coder routinely interprets rules in a different way (coder bias). However, if all the coders routinely misinterpret a coding rule, this bias will not be captured by the reliability statistic. The calculation of intercoder reliability statistics can be particularly important for research into causal interpretations or framing in media reports. Nevertheless, it is not necessarily safe to assume that hard facts are coded relatively free from errors (e.g., Eck 2012:130-35).

Lastly, it is worth noting that intercoder reliability is generally calculated at the level of the news report in communications studies. Indeed, this level allows for the closest monitoring of coder work. However, when coders are instructed to aggregate event reports and information, this monitoring process can become more complicated. Key challenges can arise when attempting to retrace coder decisions: For example, did coders notice all reports of an event, are all reports indeed about the same event, have all reports been

processed consistently, and so on. Hence, aggregation by coders, without the separate coding of news reports, can make it difficult to establish the source of low intercoder agreement in event inclusion and coding.

Nonmedia Data Comparison

To investigate errors arising from media preferences, several researchers have compared event data with nonmedia data sources. Although such data and comparisons are rare, they can give important indications of media errors. However, the external data themselves may have significant (and unknown) errors, which can jeopardize the validity of findings from media comparisons.

Police records are most frequently used to investigate the media coverage of protest events in Western contexts. Jenkins and Maher (2016:44) note that studies generally find a single newspaper covers no more (and often less) than 20–40 percent of events identified in police records. While many studies have used police records to investigate coverage error (confirming the selection effects discussed in News Coverage section), we noted that they have also been used to study reporting error. We refer in particular to the study of McCarthy et al. (1999:117-26) with regard to "hard facts" about demonstrations in Washington, DC (see News Reporting section).

Caution is nevertheless needed to avoid overly relying on the quality of police records, as they are not necessarily collected systematically and can lack important details of events (Oliver and Myers 1999:48). For events in developing contexts—the geographical focus of many conflict event data sets—police records might be subject to more serious errors than in Western contexts as well as having access problems (e.g., Bocquier and Maupeu 2005:332).

For studies focusing on armed conflict or violence against civilians, NGO reports are another external source and are commonly used for the construction of conflict event data sets (often in addition to media data). As Davenport and Ball (2002) show for state violence in Guatemala, NGO reports document more state violations and different trends in state violence over time than newspaper accounts, although whether this is due to measurement or representation error cannot be established. Interestingly, interview data show yet another picture. Further, although the purpose of many NGOs in the field is to provide independent, reliable information, reporting can be dependent on donor attention to "hot topic" events or deliberately created to draw international media attention. In turn, NGO reports often rely on media reports. Hence, NGO reports could potentially reinforce media bias toward

particular countries or conflicts in event data sets. While a military data set can reveal important insights into the coverage of armed conflict (Weidmann 2015, 2016), it can also serve particular organizational goals and does not necessarily provide a true reflection of reality.

Data Adjustments

Data weighting. The last step in the event research process is the analysis stage, during which researchers can apply corrections to the event data set to compensate for sampling or measurement error. A first type of correction involves weighting the data to correct for underrepresentation of specific events. Although the intention is to reduce error, this type of correction can also create it. Indeed, there is often no external data that match the media-based event data set. Corrections are then made based on different studies, and these findings are assumed to hold over space and time. Hug and Wisler (1998), for example, propose statistical corrections for selection bias (e.g., weighting) based on a comparative study of police records and local newspapers from four Swiss cities. They argue that corrections for coverage preferences of violent events and events with more participants might also be useful in other contexts. Ortiz et al. (2005:408-11) argue to the contrary that this can be a bias-increasing procedure if the relevance of selection factors as well as their magnitude does not translate to other contexts.

Other recently proposed corrections do not rely on comparisons with external sources but rather with other media data. Hendrix and Salehyan (2015) use a mark and recapture method to estimate the true number of events based on information from multiple media sources. SCAD draws on Associated Press (AP) and Agence France-Presse (AFP) reports. The coding scheme, starting from 2012, records whether an event was reported in AFP, AP, or both sources. By estimating the correspondence between the sources, it is possible to make corrections to the data for events not covered in both data sets. A similar approach is used proposed by Cook et al. (2017). Importantly, the method requires data sets to consistently report all sources that have reported on an event, which is not common practice. Indeed, while data sets often cite a particular source, this does not imply that the event was not included in other sources. SCAD is a notable exception. However, it does rely on the same types of media sources, international news wire services, while local newspapers could capture a substantial number of additional events (see Data Coverage section). To correct for differential attention toward particular countries by international news media (e.g., Herkenrath and Knoll 2011), it has also been proposed to include a variable for the total

number of nonconflict related news reports devoted to a particular country in a given year as a control variable in substantive analyses (Hendrix and Salehyan 2017:1664-65).

Missing data imputation. A second type of correction that can be made to conflict event data is the imputation of missing data to compensate for measurement error. These adjustments can in turn lead to erroneous statistics. Missing data corrections are often performed for dates, geolocations, and fatality estimates. For example, UCDP GED gives a date and a time to each event, but for some events, uncertainty arises about the precision of these variables (Croicu and Sundberg 2016:5-6). Sometimes only the week, month, or year of an event is known. In these cases, UCDP GED accords the earliest possible date to the event. It is also common to give the geographical coordinates of the center of the administrative unit or country when exact locations are unknown. Imputation of time and location data is often accompanied by variables indicating a level of uncertainty in the coding. Similar approaches are taken by ACLED (Raleigh and Dowd 2017:14-16). Importantly, it is not clear to what extent precision indicators are actually used in empirical applications of conflict event data, for example, by excluding uncertain events as a robustness check.

Lastly, imputations for fatality estimates also exist. One example is the splitting of the casualty count when an event occurred at multiple locations or over the course of multiple dates (e.g., ACLED but not SCAD). Another example relates to words being used to describe casualty numbers, which is relatively common (e.g., several, some, dozens). Chojnacki et al. (2012:391-92) choose to write the word down in the data set but not to quantify it. SCAD (Codebook 3.1., updated November 20, 2017:5) makes use of a distinction for missing but more ("probably large") or less ("probably small") than 10. ACLED (Raleigh and Dowd 2017:20) chooses to quantify the description: Several, many, plural, or unknown is set to 10, dozens is set to 12, hundreds is set to 100. Such quantification could potentially risk jeopardizing data quality.

Event Data: A Way Forward

The TEE framework outlined in the previous section has allowed for a comprehensive discussion of the sources of error that can affect the quality of conflict event data, cutting across subdisciplines of specialization. We have also discussed potential strategies to mitigate these errors proposed in the literature as well as their limits. Table 1 offers an

Table 1. Total Event Error: Errors, Guidelines, and Future Research.

Research Step	Error Risks	Available Estimates[a]	Solutions	Future research Directions
News source sampling	Representation and Measurement: News coverage and reporting are dependent on: – Characteristics of event: more probable selection of violent events, events with a higher number of participants/information of protests easier to establish than armed conflict events (e.g., location) – Characteristics news source: ideology, geographical scale of target audience (local, national, international), source preference (e.g., government sources) – Characteristics contexts: for example, poor infrastructure for access to (or verification of) information, government control on information	Representation: – National news reports versus police records[b]: 20–60 percent of protest events – National news reports record around 10 times less lethal violence than nongovernmental organization (NGO) documentary sources and around 2 times less than interviews[c] – International news reports versus military data[d]: 28.5 percent of lethal events – International versus national news reports[e]: 1.5–5 percent of protest/ riots; around 25 percent of lethal events; around 0.7 times the number of terrorist events – National versus provincial news reports[f]: around 0.3 times the number of deaths Measurement: – Protest report correspondence with police records[g]: >.98 (date), >.65 (purpose), >.61 (size) – Armed conflict report correspondence with military data[h]: ~80 percent within 50 km of real location; 50 percent correspondence casualties, small differences	– Draw data from multiple news sources, including media sources (with different political orientations) as well as external sources (e.g., NGO reports) – Use national or local news sources for single-country studies. International sources can be necessary for large-scale cross-national studies but hold representational risks – Adapt the scale of conflict events studied to the scale of the news source. Local sources can be better suited to study low-level events (e.g., protests versus terrorist attacks) and to conduct subnational analyses – Apply corrective weights to compensate for coverage effects based on comparisons between media sources or with external data (e.g., police records) – Control for the amount of non-conflict-related reports to account for differences in media attention toward countries or regions – Explicitly code differences between sources and take these into account for substantive analyses to account for reporting error	– Which conflict events are more likely to be selected into the news? How do biases differ according to the type of event (e.g., protest versus violent conflict)? – How do coverage and reporting differ between local, national, and international news sources? – How can we use automated coding procedures on local news sources taking into account access to digital data, language differences etc.? – How does political orientation of the news source affect event coverage and reporting? – How does press freedom affect event coverage and reporting?

(continued)

Table 1. (continued)

Research Step	Error Risks	Available Estimates[a]	Solutions	Future research Directions
News report sampling	Representation: – Issue sampling (e.g., Monday issues) and page sampling (e.g., first page) increase unreliability and can reinforce coverage bias Measurement: – Reports can rely on biased sources (e.g., government, rebels, political parties) – Reports have missing information on date, location, actors, casualties – Reports explicitly cast doubt on event occurrence, actor identity, fatality estimates	– 11–18 percent of events without precise time, 23–49 percent without precise location[j] – Uncertainty indicators make little difference[j]	– Code all issues and pages for a (random) subset of the data and compare with reduced sample. Potentially apply corrective weights – Include indicators for unreliability and bias in the codebook for event occurrence, actor identities, fatality estimates etc. – Code reports on the same event separately and include an ambiguity range in the final data set or allow users to test different specifications (e.g., minimum and maximum fatalities)	– What are the effects of different report sampling mechanisms on event statistics (e.g., front page sampling, full issue sampling, search string in online repository, automated procedure)? – How does the inclusion of unreliability and bias indicators affect substantive findings in conflict analyses?
Codebook development	Representation/Measurement – Unclear and/or ambiguous selection/coding instructions can create inconsistencies in the coding process and unsystematic as well as systematic differences between coders – Dictionaries for automated coding can leave too many relevant events out and too many irrelevant ones in and code variables incorrectly	– Sampling of a relevant event[k]: 80–97 percent recall, 23–58 percent precision – Correct categorization of event[l]: 7–96 percent	– Calculate intercoder selection/ reliability measures between human coders in a pilot phase and throughout the coding process to adapt instructions where needed – Refine the dictionary for automated procedures after data checks and rerun the coding program	– How do codebook instructions affect intercoder selection/coding reliability? – Which types of instructions work best given the reality of the data (e.g., vague reports)? – What level of complexity can we reach with human and automated coding? – What are the possibilities to improve and extend automated coding to other languages, contexts etc.?

Table 1. (continued)

Research Step	Error Risks	Available Estimates[a]	Solutions	Future research Directions
Coding process	Representation: – Coders miss relevant events or include irrelevant events – Causes unreliability but also potential bias when coders tend to miss specific events more often (e.g., low-profile protests) Measurement: – Coders wrongly categorize events or actors, make mistakes in time and location data etc. – Can cause unreliability or bias when coders systematically misinterpret instructions	– 60–70 percent overlap between independent coders in the selection of protest events[m] – Cohen's κs show substantial agreement for event category, actors, fatalities, cause[n] – Krippendorff αs for event category, actors, fatalities <0.77[o]	– Monitoring of the coding process – Calculate intercoder selection/reliability measures in a pilot phase and throughout the coding process – Retain coders who spot the most relevant events/have higher reliability scores after an initial test phase – Have a separately trained research team preselect relevant events to be coded by a different team	– What are intercoder selection/reliability measures for currently well-known data sets? – Are selection reliabilities higher when using search strings or semiautomated procedures? – How does selection interact with the characteristics of the event (e.g., violent) and the news source (e.g., common usage of visualizations)? – How reliable is the coding of hard facts (e.g., date, location, actor) versus soft facts (e.g., ascribed cause of the event)? – What are the characteristics of "good coders" (e.g., education level, length of contract)?
Nonmedia data comparison	Representation and Measurement: – External data (e.g., police records, military data, NGO reports) can also suffer from bias and unreliability – When used to correct media-based data, errors in external data can create additional uncertainty or reinforce biases	– NGO reports contain about 4 times more deaths than interview data[p]	– Compare different nonmedia sources (e.g., NGO reports, interviews, surveys) to investigate source-specific coverage and reporting effects – Drawn on media as well as external sources, code reports separately and include indicators of potential uncertainty and bias in the final data set	– What choices do writers of NGO reports make and how can this affect event data (e.g., are they likely to exaggerate atrocities to advocate for more support)? – To what extent can we rely on police records/military data in developing contexts?

Table 1. (continued)

Research Step	Error Risks	Available Estimates[a]	Solutions	Future research Directions
Data adjustments	Representation: – Corrective weights to compensate for selection effects can induce bias when selection bias is not constant over time and space. Measurement: – Imputation of missing data on dates, locations, fatalities, etc., can lead to a false sense of reliability – Time and data imputation can affect in particular analyses requiring fine-grained time and location data	– Bias in significance and direction of regression coefficients[q]	– Create weights in the data set, ensure transparency on their creation, and allow users to incorporate weights or not – Include indicators for imputation of variables	– How do corrective measures affect substantive findings? – How does selection bias differ over time, space, and news source? – How does data imputation affect substantive findings?

Note: ACLED = Armed Conflict Location and Event Dataset.

[a]Percentages are used when the events/fatalities were matched, otherwise we calculate the difference in number of events/fatalities registered (X times less or more). [b]Jenkins and Maher (2016); Myers and Canigla (2004). [c]Davenport and Ball (2002). [d]Weidmann (2016). [e]Demarest and Langer (2018); Herkenrath and Knoll (2011), Bueno de Mesquita et al. (2015). [f]Barron and Sharpe (2008). [g]McCarthy et al. (1999), print media estimates. [h]Weidmann (2015). [i]ACLED and UCDP data, respectively. [j]Schrodt and Ulfelder (2016). [k]Bond et al. (1997), Croicu and Weidmann (2016), King and Lowe (2003). [l]Bond et al. (1997), Boschee et al. (2013), King and Lowe (2003), Stepinski et al. (2006). [m]Hutter (2014), Kriesi et al. (1998). [n]Salehyan et al. (2012). [o]Demarest and Langer (2018). [p]Davenport and Ball (2002). [q]Ortiz et al. (2005).

25

overview of these error sources, available estimates of their size, and mitigation strategies. The estimates of the degree of error are based on the studies reviewed here and hence on different geographical contexts, time periods, (automated) coding procedures, and so on. Moreover, the estimates show the extent to which information can diverge but not necessarily how this impacts substantive research results. While this should be taken into account, they do provide researchers indications on how to assess data quality. Finally, the available (and unavailable) estimates also indicate where more empirical research is needed. In this section, we focus mostly on the methodological questions which have so far been insufficiently addressed in the literature. The last column of Table 1 contains an extensive list of questions which require further research and which together constitute a research agenda concerning the methodology of creating and using conflict event data sets.

Most attention in the literature has been directed to coverage error and for important reasons. Indeed, the available estimates on event selection reveal that the distorting effects of coverage error on research findings may be substantial. While most evidence of such bias has been established in the context of protest movements in Western contexts (e.g., Jenkins and Maher 2016), there is indication that violent conflict as well is underreported (Davenport and Ball 2002; Weidmann 2016). Besides the form of conflict, another important challenge concerns the widespread use of international news wire reports to investigate (violent) conflict in the developing world. Evidence suggests that this may be problematic (Bueno de Mesquita et al. 2015; Demarest and Langer 2018; Herkenrath and Knoll 2011). This problem could be mitigated by the increased availability of online local sources and, potentially, new evolutions in automated coding. Interestingly, the available estimates on reporting error appear to indicate that the facts of protests (McCarthy et al. 1999) and violent conflict (Weidmann 2015) may be reported relatively error free. Nevertheless, reporting error can also be dependent on the context. This is especially important to take into account when considering local media sources subjected to government control. No estimates appear to be available for these types of contexts, however.

Errors arising from the logic of the media source require careful consideration and further research. Other features of the data collection protocol require attention too, however. This is also revealed by the estimates of errors related to the coding process. In this regard, it is important to point out that while different indicators can be used to assess particular methodological choices (selection agreement, intercoder reliability, recall, and precision), there is for now no real consensus in the

literature concerning the use of such indicators and, consequently, their reporting. Many new data sets in peace and conflict studies for instance rarely provide information on coder selection and reliability or the general degree of imputed data in the data set. By contrast, automated coding developers appear to show more agreement on the need to report recall and precision rates.

The measurement of such errors is important to establish where most data collection efforts should be directed in order to achieve the largest gains in terms of data quality. For example, Schrodt and Ulfelder (2016:29) argue that including indicators of uncertainty about events, actors, and so on (see Table 1, "codebook development"), did not add much value to PITF's atrocities data, and they left out these indicators in later versions. The same questions can be raised with regard to the coding of reports separately to account for differences between them (Day et al. 2015; Weidmann and Rød 2015). More research is needed in order to determine the merits of such procedures to be able to assess their use for new data sets.

Users as well should direct sufficient attention to event error sources. This applies to the selection of particular data sets to address substantive research questions but also in reporting and robustness checks. Event error sources are necessary to understand the limits of particular studies, for instance, both in the academic and policy domains. Furthermore, when data developers provide indicators of data quality, we argue that researchers focusing on substantive questions should not only report these indicators in their studies but should also reflect upon the implications of these indicators for the validity of their findings and conclusions. This includes in particular missing data imputation indicators that are not commonly used in quantitative conflict studies even though the field is increasingly focusing on fine-grained details on events both in time and in place (Gleditsch et al. 2014). Although event data weighting is still not commonly used, again the effect of weighting should be carefully compared with results based on nonweighted data, and a preference for some results over others should be explicitly motivated.

Conclusion

The quality of conflict event data can be affected by a wide range of errors. The discussion in this article was guided by the current state of the art concerning conflict event studies and also drew on social movement and communications studies. The major advantages of the TEE framework is

that it offers a holistic perspective on the sources of error affecting conflict event data and, by consequence, analytical clarity into an arguably broad field of study. Indeed, while many error sources have been discussed in the literature, these debates have not always allowed for further systematization. By doing just this, the TEE framework offers a baseline tool for new and established data developers and users, as well as guidance for future research. Furthermore, while TEE has focused on human and automated event data collection practices, it can be extended into new areas. The emergence of "citizen reporting" via social media, for example, is becoming an important new source for event data collection but similar concerns with regard to coverage and reporting effects, as well as data collection procedures apply.

Finally, it is worth noting that while errors can and should be minimized, they can hardly be ruled out completely. Hence, event data will never be a true reflection of reality. However, this is not unlike other empirical data sources in the social sciences, including public opinion surveys. Going back to our initial analogy, it is worth considering that the sources of error are widely recognized in survey research but also that the real exercise lies in minimizing errors by taking into account limited resources. As with "survey errors and survey costs" (Groves 1989), the balance between event errors and costs constrains event data set developers. In order to improve guidelines and standards for data collection, however, conflict event data methodology needs to be considered as a research agenda in its own right. The TEE framework and the research questions laid out in Table 1 offer important directions to do so.

ORCID iD

Leila Demarest https://orcid.org/0000-0001-6887-9937

Notes

1. Earlier versions of UCDP Georeferenced Event Dataset included only conflict events in Africa; a new global data set is currently available (version 5.0). Note

that especially for violent conflict event data sets, developing countries are predominant, even if the data set has a global focus.

2. An extensive range of studies has been written on the topic of media selection bias alone. We provide an overview of key ideas here and direct readers to the references cited in this section, and Data Weighting section for corrections on selection bias, for further information.

3. All the conflict event data sets mentioned in the introduction predominantly use these inventories, except for Armed Conflict Location and Event Dataset, which also draws from local newspapers.

4. The use of local news sources has, however, been more frequently the case to investigate Hindu–Muslim riots in India (e.g., Wilkinson 2004).

5. Note that in literature, hard news is also conceptualized as having a high degree of newsworthiness such as news regarding politics, economics, and social matters, whereas soft news has less substantive informational value, for example, gossip, human interest stories, and so on (e.g., Reinemann et al. 2011). This distinction is related to, but differs from, the one used in this article.

6. While the dictionary and the coding program are in principle separate entities in automated coding procedures (Schrodt and Van Brackle 2013:24), we do not explicitly separate the two in our discussion here. We also do not go into programming errors or characteristics (e.g., speed).

7. Recall is equal to the number of true positives divided by the sum of the number of true positives and the number of false negatives. Precision is equal to the number of true positives divided by the sum of the number of true positives and the number of false positives. The F1 statistic captures the harmonic mean of recall and precision (Heap et al. 2017).

8. The sparse parsing procedure breaks down sentences in relevant text based on actors, targets, and transient verbs.

9. It is, however, important to mention that the use of TABARI in their work has been criticized by Schrodt (Schrodt and Van Brackle 2013:27).

10. See Codebook 3.1, updated November 20, 2014, pp. 3-4.

11. In addition to intercoder reliability one can pay attention to intracoder reliability or stability, that is, does a coder code previous reports in the same way at a later point in time (Krippendorff 2013:270-71).

References

Azar, Edward E. 1980. "The Conflict and Peace Data Bank (COPDAB) Project." *Journal of Conflict Resolution* 24:143-52.

Barron, Patrick and Joanne Sharpe. 2008. "Local Conflict in Post-Suharto Indonesia: Understanding Variations in Violence Levels and Forms through Local Newspapers." *Journal of East Asian Studies* 8:395-423.

Bernauer, Thomas and Nils P. Gleditsch. 2012. "New Event Data in Conflict Research." *International Interactions* 38:375-81.

Bocquier, Philippe and Hervé Maupeu. 2005. "Analysing Low Intensity Conflict in Africa Using Press Reports." *European Journal of Population* 21:321-45.

Bond, Doug, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor. 2003. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development." *Journal of Peace Research* 40:733-45.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor, and Kurt Schock. 1997. "Mapping Mass Political Conflict and Civil Society: Issues and Prospects for the Automated Development of Event Data." *Journal of Conflict Resolution* 41:553-79.

Boschee, Elizabeth, Premkumar Natarajan, and Ralph Weischedel. 2013. "Automatic Extraction of Events from Open Source Text for Predictive Forecasting." Pp. 51-67 in *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian. New York: Springer Science + Business Media.

Bueno de Mesquita, Ethan, C. Christine Fair, Jenna Jordan, Rasul B. Rais, and Jacob N. Shapiro. 2015. "Measuring Political Violence in Pakistan: Insights from the BFRS Dataset." *Conflict Management and Peace Science* 32:536-58.

Chan, Joseph M. and Chi-Chuan Lee. 1984. "Journalistic 'Paradigms' of Civil Protests: A Case Study in Hong Kong." Pp. 249-76 in *The News Media in National and International Conflict*, edited by Andrew Arno and Wimal Dissanayake. Boulder, CO: Westview.

Chojnacki, Sven, Christian Ickler, Michael Spies, and John Wiesel. 2012. "Event Data on Armed Conflict and Security: New Perspectives, Old Challenges, and Some Solutions." *International Interactions* 38:382-401.

Collier, Paul and Anke Hoeffler. 2002. "Greed and Grievance in Civil War." Centre for the Study of African Economies Working Paper Series 2002-01, Oxford, England.

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* 25:223-40.

Croicu, Mihai and Ralph Sundberg. 2016. *UCDP GED Codebook Version 5.0.* Uppsala, Sweden: Department of Peace and Conflict Research, Uppsala University.

Croicu, Mihai and Nils B. Weidmann. 2015. "Improving the Selection of News Reports for Event Coding Using Ensemble Classification." *Research & Politics* 2:1-8.

Dardis, Frank E. 2006. "Marginalization Devices in U.S. Press Coverage of Iraq War Protest: A Content Analysis." *Mass Communication and Society* 9:117-35.

Danzger, M. Herbert. 1975. "Validating Conflict Data." *American Sociological Review* 40:570-84.

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

Davenport, Christian and Patrick Ball. 2002. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46:427-50.

Day, Joel, Jonathan Pinckney, and Erica Chenoweth. 2015. "Collecting Data on Nonviolent Action: Lessons Learned and Ways Forward." *Journal of Peace Research* 52:129-33.

Demarest, Leila and Arnim Langer. 2018. "The Study of Violence and Social Unrest in Africa: A Comparative Analysis of Three Conflict Event Datasets." *African Affairs* 117:310-25.

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.

Eck, Kristine. 2012. "In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Event Datasets." *Cooperation and Conflict* 47:124-41.

Eisinger, Peter K. 1973. "The Conditions of Protest Behavior in American Cities." *The American Political Science Review* 67:11-28.

Entman, Robert M. 1993. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication* 43:51-58.

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97:75-90.

Franzosi, Roberto. 1987. "The Press as a Source of Socio-historical Data: Issues in the Methodology of Data Collection from Newspapers." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 20:5-16.

Galtung, Johan and Mari H. Ruge. 1965. "The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers." *Journal of Peace Research* 2:64-90.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Hoboken, NJ: Wiley.

Gleditsch, Kristian S., Nils W. Metternich, and Andrea Ruggeri. 2014. "Data and Progress in Peace and Conflict Research." *Journal of Peace Research* 51:301-14.

Hammond, Jesse and Nils B. Weidmann. 2014. "Using Machine-coded Event Data for the Micro-level Study of Political Violence." *Research & Politics* 1:1-8.

Harcup, Tony and Deirdre O'Neill. 2001. "What Is News? Galtung and Ruge Revisited." *Journalism Studies* 2:261-80.

Harcup, Tony and Deirdre O'Neill. 2016. "What Is News? News Values Revisited (Again)." *Journalism Studies* 18:1-19. doi: 10.1080/1461670X.2016.1150193.

Heap, Bradford, Alfred Krzywicki, Susanne Schmeidl, Wayne Wobcke, and Michael Bain. 2017. "A Joint Human/Machine Process for Coding Events and Conflict Drivers." Conference Paper, International Conference on Advanced Data Mining and Applications, ADMA 2017, Singapore. October.

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* 41:392-406.

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

Herkenrath, Mark and Alex Knoll. 2011. "Protest Events in International Press Coverage: An Empirical Critique of Cross-national Conflict Databases." *International Journal of Comparative Sociology* 52:163-80.

Hickler, Christian and John Wiesel. 2012. "New Method, Different War? Evaluating Supervised Machine Learning by Coding Armed Conflict." SFB-Governance Working Paper Series, No.39. Collaborative Research Center (SFB) 700, Berlin. 2012. Retrieved October 12, 2019 (https://refubium.fu-berlin.de/bitstream/handle/fub188/18738/WP39.pdf?sequence=1&isAllowed=y).

Hirshleifer, Jack. 1994. "The Dark Side of the Force." *Economic Inquiry* 32:1-10.

Hug, Simon and Dominique Wisler. 1998. "Correcting for Selection Bias in Social Movement Research." *Mobilization: An International Quarterly* 3:141-61.

Hutter, Sven. 2014. "Protest Event Analysis and Its Offspring." Pp. 335-67 in *Methodological Practices in Social Movement Research*, edited by Donatella della Porta. Oxford, UK: Oxford University Press.

Jenkins, J. Craig and Thomas V. Maher. 2016. "What Should We Do about Source Selection in Event Data? Challenges, Progress, and Possible Solutions." *International Journal of Sociology* 46:42-57.

Kalyvas, Stathis N. 2006. *The Logic of Violence in Civil War*. Cambridge, MA: Cambridge University Press.

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57:617-42.

Koopmans, Ruud and Dieter Rucht. 2002. "Protest Event Analysis." Pp. 231-59 in *Methods of Social Movement Research*, edited by B. Klandermans and S. Staggenborg. Minneapolis: University of Minnesota.

Kriesi, Hanspeter, Ruud Koopmans, Jan W. Duyvendak, and Marco G. Giugni. 1998. *New Social Movements in Western Europe: A Comparative Analysis*. Minneapolis: University of Minnesota.

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.

Leetaru, Kalev and Philip A. Schrodt. 2013. "GDELT: Global Data on Events, Location and Tone, 1979-2012." Paper presented at the International Studies Association Meetings, San Francisco, CA, April.

LaFree, Gary and Laura Dugan. 2007. "Introducing the Global Terrorism Database." *Terrorism and Political Violence* 19:181-204.

Lee, Francis L. F. 2014. "Triggering the Protest Paradigm: Examining Factors Affecting News Coverage of Protests." *International Journal of Communication* 8: 2725-46.

McCarthy, John D., Clark McPhail, Jackie Smith, and Louis J. Crishock. 1999. "Electronic and Print Media Representations of Washington D.C. Demonstrations, 1982 and 1991: A Demography of Description Bias." Pp. 113-30 in *Acts of Dissent: New Developments in the Study of Protest*, edited by Dieter Rucht, Ruud Koopmans, and Friedhelm Neidhart. Oxford, England: Rowman & Littlefield.

McClelland, Charles A. 1976. *World Event/Interaction Survey Codebook (ICPSR 5211)*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

McLeod, Douglas and James Hertog. 1992. "The Manufacture of 'Public Opinion' by Reporters: Informal Cues for Public Perceptions of Protest Groups." *Discourse & Society* 3:259-75.

Myers, Daniel J. and Beth S. Caniglia. 2004. "All the Rioting That's Fit to Print: Selection Effects in National Newspaper Coverage of Civil Disorders, 1968-1969." *American Sociological Review* 69:519-43.

Oliver, Pamela E. and Daniel J. Myers. 1999. "How Events Enter the Public Sphere: Conflict, Location, and Sponsorship in Local Newspaper Coverage of Public Events." *American Journal of Sociology* 105:38-87.

Ortiz, David G., Daniel J. Myers, N. Eugene Walls, and Maria-Elena D. Diaz. 2005. "Where Do We Stand with Newspaper Data?" *Mobilization: An International Journal* 10:397-419.

Raleigh, Clionadh and Caitriona Dowd. 2017. "Armed Conflict and Event Location (ACLED) Codebook." Retrieved 12 October 2019 (https://www.acleddata.com/wp-content/uploads/2017/01/ACLED_Codebook_2017.pdf).

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED-Armed Conflict Location and Event Data." *Journal of Peace Research* 47:1-10.

Reinemann, Carsten, James Stanyer, Sebastian Scherr, and Guido Legnante. 2011. "Hard and Soft News: A Review of Concepts, Operationalizations and Key Findings." *Journalism* 13:221-39.

Ruggeri, Andrea, Theodora-Ismene Gizelis, and Han Dorussen. 2011. "Events Data as Bismarck's Sausages? Intercoder Reliability, Coders' Selection, and Data Quality." *International Interactions* 37:340-61.

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

Salehyan, Idean, Cullen S. Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. "Social Conflict in Africa: A New Database." *International Interactions* 38:503-11.

Schrodt, Philip A. 2006. "Twenty Years of the Kansas Event Data System Project." *The Political Methodologist* 14:2-8.

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

Schrodt, Philip A. and David Van Brackle. 2013. "Automated Coding of Political Event Data." Pp. 23-49 in *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian. New York: Springer Science + Business Media.

Schrodt, Philip A., Erin M. Simpson, and Deborah J. Gerner. 2001. "Monitoring Conflict Using Automated Coding of Newswire Reports: A Comparison of Five Geographical Regions." Conference paper, June 8-9, Uppsala, Sweden.

Schrodt, Philip A. and Jay Ulfelder. 2016. "Political Instability Task Force Atrocities Event Data Collection Codebook Version 1.1b1." Retrieved October 12, 2019 (http://eventdata.parusanalytics.com/data.dir/PITF_Atrocities.codebook.1.1B1.pdf).

Smith, Jackie, John D. McCarthy, Clark McPhail, and Augustyn Boguslaw. 2001. "From Protest to Agenda Building: Description Bias in Media Coverage of Protest Events in Washington, D.C." *Social Forces* 79:1397-423.

Schneider, Gerald and Margit Bussmann. 2013. "Accounting for the dynamics of one-sided violence: Introducing KOSVED." *Journal of Peace Research* 50:635-44.

Stepinski, Adam, Richard Stoll, and Devika Subramanian. 2006. "Automated Event Coding Using Conditional Random Fields." Retrieved June 13, 2018 (https://www.cs.rice.edu/~devika/conflict/papers/draft1.pdf).

Sundberg, Ralph and Erik Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50:523-32.

Urdal, Henrik. 2008. "Urban Social Disturbance in Africa and Asia: report on a New Dataset." PRIO Papers, Oslo, Norway.

Weaver, David A. and Joshua M. Scacco. 2012. "Revisiting the Protest Paradigm." *The International Journal of Press/Politics* 18:61-84.

Weidmann, Nils B. 2015. "On the Accuracy of Media-based Conflict Event Data." *The Journal of Conflict Resolution* 59:1129-49.

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

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

Wilkinson, Steven I. 2004. *Votes and Violence: Ethnic Competition and Ethnic Riots in India*. New York: Cambridge University Press.

Author Biographies

Leila Demarest is an assistant professor of African politics at the Department of Political Science, Leiden University, the Netherlands. Her research interests include social movements and political mobilization in Africa, political communication, and quantitative and qualitative social science research methodology. Her most recent publications (with Arnim Langer) are "The Study of Violence and Social Unrest in Africa: A Comparative Analysis of Three Conflict Event Datasets" in *African Affairs* (April 2018) and "Peace Journalism on a Shoestring? Conflict Reporting in Nigeria's National News Media" in *Journalism* (first online, August 2018).

Arnim Langer is a professor in international relations at KU Leuven and director of the Centre for Research on Peace and Development (CRPD) at the Faculty of Social Sciences. Currently, he is also Humboldt Research Fellow at the University of Heidelberg, Germany. He has published extensively on the causes of violent conflict in heterogeneous societies and the challenges to sustainable peacebuilding. Some of his recent publications include "Conceptualising and Measuring Social Cohesion in Africa: Towards a Perceptions-based Index" (published in *Social Indicators Research*) and "A General Class of Social Distance Measures" (published in *Political Analysis*).