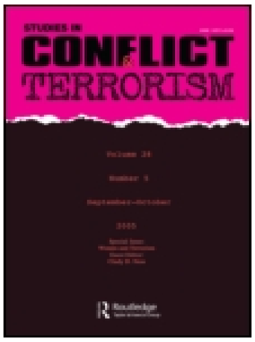

Routledge
Taylor & Francis Group

Studies in Conflict & Terrorism

ISSN: 1057-610X (Print) 1521-0731 (Online) Journal homepage: http://www.tandfonline.com/loi/uter20

# Speaking Truth to Sources: Introducing a Method for the Quantitative Evaluation of Open Sources in Event Data

Gary A. Ackerman & Lauren E. Pinson

**To cite this article:** Gary A. Ackerman & Lauren E. Pinson (2016) Speaking Truth to Sources: Introducing a Method for the Quantitative Evaluation of Open Sources in Event Data, Studies in Conflict & Terrorism, 39:7-8, 617-640, DOI: 10.1080/1057610X.2016.1141000

**To link to this article:** http://dx.doi.org/10.1080/1057610X.2016.1141000

Published online: 06 Apr 2016.

STUDIES IN CONFLICT & TERRORISM
2016, VOL. 39, NOS. 7–8, 617–640
http://dx.doi.org/10.1080/1057610X.2016.1141000

Routledge
Taylor & Francis Group

# Speaking Truth to Sources: Introducing a Method for the Quantitative Evaluation of Open Sources in Event Data

Gary A. Ackerman and Lauren E. Pinson

National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland, College Park, MD, USA

**ABSTRACT**

Open-source event data sets frequently used for social science analysis rarely provide any transparent explanation of the credibility of sources or the validity of data thereby obtained. We develop a sample Source Evaluation Schema for the purpose of operationalizing measures of open-source event validity at the case, source, and variable levels. Based on our findings, we argue that explicitly incorporating and disclosing credibility and validity levels allows for greater flexibility in tailoring the inclusion of cases for researchers' specific analytical requirements. By facilitating more transparent analyses, the inclusion of such measures in similar datasets can result in more defensible conclusions, especially in highly charged political and security contexts such as those surrounding terrorism.



> When we hear news we should always wait for the sacrament of confirmation.
>
> —Voltaire

> It's hard to tell if the world is actually growing worse, or if the news coverage is just better.
>
> —Joe Moore, U.S. television personality[1]

The analysis of empirical data in the pursuit of greater understanding is a bedrock of the social sciences. In many circumstances, first-hand collection of such data is impossible or impractical given resource constraints, and researchers must rely on publicly available information collected and recorded by others, including journalists, eyewitnesses, and court clerks. Moreover, a decade or more after an event, news articles, government reports, or court documents are often the only sources available. These "open sources" have proven to be an invaluable resource to various types of social scientists, and have recently even been accorded significance by the traditionally hyper-parochial intelligence community.[2] Yet, deriving sound insights requires that the data used in analysis be reliable, or at the very least that the reliability of the data be taken into consideration during analysis. With renewed interest in evidence-based policy[3] and growing opportunities for the social sciences to bring rigorous analysis to the policymaking process, the validity of open-source data that is applied

**CONTACT** Gary A. Ackerman gackerman@start.umd.edu National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland, College Park, MD, USA.

Color versions of one or more figures in the article can be found online at www.tandfonline.com/uter.

to inform policy becomes critical. Otherwise, no matter how rigorous the analysis or sincere the intentions of researchers and decision makers might be, the result could be ineffective or deleterious policy. This is especially true as it relates to influencing policy surrounding pressing and contentious social issues, such as those relating to terrorism.

All of this may seem trite and deserving of little more than inclusion in an introductory research methods textbook. This is why it might at first seem surprising that despite the rather pedestrian conceptual issues involved and a body of scholarly discussion going back at least fifty years, major event data sets—especially those related to international security—seldom explicitly incorporate the credibility of open sources or the validity of their data. In light of this absence of explicit validation measures and the potential policy relevance thereof, this article aims to revive this debate, at the same time recognizing the abundance of open sources and the proliferation of new media prevalent in the Information Age.

Admittedly, establishing a standard source evaluation schema that can be implemented in large-scale databases with minimal resources requires an interdisciplinary discussion. The current effort was driven, however, from a more particular need on the part of the authors—the creation of an event-level data set of chemical, biological, radiological, and nuclear (CBRN) terrorism-related activities. We were aware from the outset that analysis of this data set would be used to inform national security policy in this area and that many of the incidents reported in the open sources were controversial. We thus endeavored to ensure, or at least to quantify, our level of confidence that the events we were recording actually occurred as reported, since the inclusion of apocryphal cases or incorrect details could seriously distort our subsequent analysis.[4] We were initially astonished when we began examining the codebooks and methodology documents of similar data sets for guidance in this regard and found that such measures were scarcely mentioned, let alone systematically and rigorously implemented. This prompted us to explore the issue more deeply and eventually to develop the Source Evaluation Schema described below.

Although designed for our particular research enterprise of terrorism data collection, we believe this approach has wider application and would strengthen similar social science data collection and coding efforts, in particular those relating to politically charged or controversial topics, where various incentives might exist to distort the reporting of events.

Particularly in the domain of terrorism, official reporting is often insufficient for scientific analysis. First, many countries and localities might lack the resources to properly collect and record statistics on terrorist events occurring within their jurisdiction, resulting in the lack of official data in many areas of the world. Second, even where these resources exist, many governments, especially those of the less democratic ilk, might find it politically expedient to obfuscate the incidence of terrorism to serve the regime's interests, either censoring terrorism data that the government fears could undermine its legitimacy, or alternatively inflating the rate of terrorism in order to provide a scapegoat or justification for pursuing particular enemies. Third, even in those cases where official sources do not consciously dissemble, governments adopt varying definitions of terrorism that reflect parochial concerns, making it problematic, especially for global data sets, to obtain consistent data from official reports. One of the strengths of open sources is that they can offer multiple, often alternative viewpoints beyond those of official reports or mainstream media, which can be compared with each other and with official data. Yet, the very variety in the provenance and orientation of these open sources means that their advantages can only be fully realized if there is some basis for rating the reliability of the sources and the information they provide. Most importantly, any such rating needs to be explicitly encoded in the data itself if it is to be useful to analysis.

We thus present our schema, together with an illustrative demonstration of its utility using the data set we developed. This is intended, at the very least, to serve as a catalyst for further refinement and discourse. We begin by presenting a review of existing scholarship on open source validation, followed by establishing definitions for credibility and validity and detailing the need for explicit source and information validation. Next, we discuss the (lack of) use of such measures in extant data sets. We then describe our approach and how we generated our schema before presenting a brief illustration of the use of these measures in a simple analysis.

## Determining Validity in Open Sources

To guide our choice of metrics, we initially considered the body of literature on evaluating event data in domains of study prone to varying sympathies, conflicting reports, and limited first-hand information. Such domains include human rights abuses, civil wars, crime, collective action incidents, racial violence, and terrorist incidents.[5] Next, we framed two terms, credibility and validity, aiming to counteract two of the major biases discussed in source evaluation literature. Then, we sought guidance by consulting various international relations data sets themselves. In combination, these approaches helped reveal the main problems with, and suggest best practices for, establishing credibility and validity measures in events data.



### *Selection Bias*

Snyder and Kelly's groundbreaking study on the validity of newspaper data analyzed the probability that events will be reported relative to event intensity and media sensitivity.[6] Notably, they argued that above a certain size event, the media reports violent events without selection bias. While this analysis advocates for addressing selection bias in newspaper data, it does not address the accuracy of published events. Nor does it speak to the possibility of variability in selection effects, as discussed by Oliver and Maney in the context of protest events, who convincingly conclude that "it is simply not possible to assert, in the absence of data, that the patterns of selection in news coverage … should be assumed to be relatively stable across time or locale of issue."[7] Reeves et al. argue that researchers may commit Type I and Type II errors due to source bias created by using a data set based on a single source. They conclude that researchers analyzing events data should compare results against multiple sources or use multisource data.[8] Chermak et al. evaluate the potential selectivity bias of the sources used to construct the United States Extremist Crime Database (ECDB); the authors argue that open source data sets (drawn from multiple sources) provide information "that is representative of the larger universe [of cases of interest]."[9]

Robustness of findings across data sources might therefore increase the confidence in any conclusions drawn,[10] but even when utilizing multiple sources, Hug and others discuss the selection bias created by incomplete data sets.[11] According to Schrodt, the field has not reached consensus on how many sources are sufficient or when local, and especially non-English, sources are essential.[12] Further, open source data sets may over represent more "newsworthy" events; for example, in the context of the Global Terrorism Database (GTD), LaFree and Dugan discuss how media sources are less likely to report on stymied plots and events that take place in locations with less media coverage.[13]

## Description Bias

Many event-centered domains of study rely on news sources as a main source of data even though media sources have inherent description biases. Earl et al. define description bias as "the veracity with which selected events are reported in the press."[14] Even normally inclusive and cautious news sources may publish biased and less than comprehensive versions of events. When events data ignores the varying perspectives derived from diverse news sources, errors in statistical analysis ranging from invalid causal inferences to measurement bias may become unavoidable. With reference to the stability of this type of bias, Davenport and Litras's analysis of news sources reporting repression against the Black Panther Party suggests that the content of coverage varied systematically. Mainstream media sources provided more coverage of authorities while sources sympathetic to the perpetrators gave more information on the dissidents.[15] On the other hand, Woolley notes that news sources have biases that usually fluctuate over time.[16] In their discussion of newspapers reporting social movements, Barranco and Wisler maintain that press bias should be neutralized in analysis.[17] They support using sources with competing biases to create a complementary analysis. In his discussion of the media's role in framing issues, Entman advocates for increased measurement of slant and bias.[18] Strawn contends that the coverage tendencies of source institutions can be defined relative to other institutions. He notes that identifying absolute biases may be excessively expensive and time-intensive. Yet, by discerning institutional tendencies in a relative fashion, researchers can better control for source biases.[19]



## Toward More Precise Formulations of Credibility and Validity

Assessing and coding the credibility of sources and the validity of data and cases can help to mitigate selection bias and description bias. Many of these terms, such as validity and credibility, are associated with traditional mathematical notions of measurement and can provide guidance in more carefully characterizing the confidence we have in open sources,[20] but only up to a point. While being guided—where this proves useful—by statistical measurement terminology, we therefore provide distinct formulations of these terms to suit our current purposes.

To begin with, the validity of a measurement formally equates to the degree to which it faithfully represents the concept in question.[21] In the context of this arrticle, we define validity—both at the level of individual data points and of the events as a whole—as *the level of confidence that the information that is recorded objectively reflects the reality of what is being measured*, or what is sometimes referred to as "ground truth." In an open-source event data set, validity is thus based on determining how closely the extant available records of the event (whether news articles, government reports, or court transcripts) represent the event as it actually occurred. This process becomes complicated, as we shall see below, by the general lack—at least on the part of most researchers—of access to a perfect record of what actually occurred against which to compare reported event details.

As for the credibility measure, we see that a general definition of credibility portrays it as "the quality or power of inspiring belief,"[22] but this proves to be too vague for our purposes.

Journalistic credibility is often measured as a multidimensional construct that may include fairness, accuracy, comprehensiveness, and believability.[23] Basically, credibility reflects the degree of confidence we have that the source we are using is an undistorted representation of what was observed. Since we are not, epistemologically speaking, absolutely certain about what exactly was observed at the time of the event, we can measure credibility relative to a hypothesized ideal source. In the context of the current enterprise, we define the *credibility* of a source as *the likelihood that an additional completely competent and disinterested source would report the same information based on the same event*. When considering open sources in an event data set, source credibility is thus derived from the general perception that a source should be dependable and impartial.[24] For practical purposes, this credibility emerges from an evaluation of both the content of the source itself and the context, usually the publishing institution and/or author, associated with the source. Highly credible sources are objective and skilled at reporting on the subject matter. They are most likely to report events authentically. These sources are often corroborated by other well-regarded independent sources. In contrast, the most non-credible sources are biased and/or unskilled and their reporting will often contrast with that of other independent sources of greater credibility.

So, whereas validity measures pertain more to how accurately reality is represented, credibility measures focus on the attributes of the reporting entity. The two concepts are, however, in practice and in theory, often difficult to isolate from one another. For practical purposes, then, it may be easier to devise a single variable that performs the desired functions of both validity and credibility evaluation—and thus implicitly links them. This can be achieved by defining a validity measure for the event as a whole that is based on an aggregation of the credibility levels of each relevant source. In the absence of access to an absolutely unimpeachable record of the ground truth, this *source-derived validity* would provide a practical measure of epistemological confidence in existing open source representations of the event.

While credibility is usually evaluated at the source level, and source-derived validity at the event level, there is a third level of measurement, namely the level of individual variable values. We refer to evaluation at this level as *data validation*. Each and every piece of information presented (e.g., actor identity, number of casualties, or type of event) may not be consistent across sources, even when these sources are more or less equally credible. Again, since there is no way to unequivocally "know" ground truth, at least for events examined second-hand, the only way to usefully measure validity in the current context is in the negative, by highlighting those instances where the sources shed some doubt on the validity of a piece of information. This is achieved practically by looking for a lack of corroboration between sources. For instance, while several sources may agree on most of the details of a particular terrorist incident, there may be no agreement across the obtained sources as to the number of victims involved. Obviously, the credibility of each source must be considered when determining the validity of information presented therein. This implies that—just as in the case of credibility and event validity—credibility and data validation are, in practical, if not conceptual, terms, usually interdependent. Moreover, confidence in certain variables within the case may be stronger than confidence in the case as a whole, or vice versa.

## The Need for Explicit Source Evaluation in Open-Source Databases

News sources are likely to be the main source of information for many event data sets. Although traditional media organizations generally have more structured reporting procedures when compared with other sources, such as online nontraditional media (blogs, social media, etc.), even traditional media sources are not always guaranteed to contain accurate information. Neither is this phenomenon limited to media reports—even seemingly authoritative sources like government reports and court transcripts are susceptible to distortions in the recording of information.

There are several reasons why—in the absence of evidence to the contrary, such as an explicit process of evaluation—one might be justified in regarding open source data found in social science databases with some degree of circumspection. The most common of these are:

- *Coding error*—"natural" human error, encompassing such factors as inexperience, cognitive deficits (fatigue, fallacious logic, intoxication, etc.), and affect-based distortions (brought about by prejudice, jealousy, etc.). These errors are often exacerbated by insufficiently specified variables or overly subjective coding requirements.
- *Limited or no source record*—the coding procedure does not require coders to explicitly record the sources used for each separate event. Without explicit source recording, coding replication is not possible.[25]
- *Lack of information*—the coder may be unable to obtain sufficient or sufficiently detailed source materials owing to either (a) a genuine dearth of published, publicly available sources of information or (b) an inability, for a variety of reasons (such as resource or language limitations), to access existing sources. For instance, while a focus on English-only sources and available translations may suffice in much of the world, not including Spanish-language stories in Latin America likely leads to selection bias.
- *Unreliable sources*—data derived from open sources may be called into question where the information contained in one or more sources has been distorted either (a) intentionally through a systematic institutional or author bias or (b) accidentally due to a lack of competence or resources.
- *Inherent uncertainty*—in some events, even where the known facts have been reported reliably and in full, the nature of the event remains indeterminate. For instance, there may not be sufficient forensic evidence left behind for authorities to be able to determine if an explosion at a chemical plant was an attack or an accident, or authorities may not conclusively determine the initial catalyst of a protest.

Appropriate codebook construction, coder training, and intercoder reliability checks can help address the first two aforementioned difficulties,[26] but to address the latter three issues one must turn to a close examination of the data sources themselves.[27]

Several suggestions regarding how to conduct source evaluation emerge from the social science literature. The literature on disputed events data supports the need for a broad range of sources to eliminate selection bias and neutralize reporting bias. Using a single source virtually ensures the creation of selection bias, although even using multiple sources is no panacea for eliminating bias overall. For instance, sources sharing the same bias are unlikely to yield a balanced view of events. Studying the entire population of cases from which the sample is drawn is one possible way to correct for selection bias, but identifying the universe of events, let alone obtaining a complete characterization or data set thereof, may be impractical in many topic areas.[28] We argue that an adequate solution, however, might be provided by creating a dataset with a wide range of information about its sources combined with a transparent measure of its variable and overall case validity.

Most authors also maintain that the credibility of the author and publisher must be evaluated.[29] In turn, they stress that the objectivity, timeliness, and corroboration of a source all play a role in analyzing its credibility. In terms of how to go about comprehensive examination of a source's credibility, the literature generally recommends relying on both intrinsic and extrinsic evaluation. Intrinsic evaluation is based on the quality of the source material itself (e.g., an individual transcript or newspaper article) and involves close analysis by the researcher of the content, style, and structure of the source document. This type of analysis can provide what has been termed the "surface credibility" of a source, which in turn affects its believability.[30] Basic examples of indicators of low surface credibility include grammatical errors, internal inconsistencies, adjoining advertisements that reflect a bias, persuasive writing that disregards facts, and a lack of adequate citation. The observed quality of a source's content thus presents one means of assessing a level of confidence in the source. Extrinsic evaluation, on the other hand, is contextual. It considers such factors as the reputation of the author and the institutional publisher. For instance, a mission statement summarizing the publication's aims may reveal a general bias likely to color most articles published therein.

Notably, not only should some level of source evaluation be conducted, but we argue that any procedures should be transparent and explicitly coded into the database itself. We can conceive of at least five reasons for *explicitly* including source credibility and validity measures in event-based data sets derived from open sources:

1. More defensible conclusions can be drawn from analysis using the data set if credibility and validity information is presented with the analysis and results are found to be insensitive to various levels of source and data reliability.
2. Based on the (often mistaken) assumption that a case's inclusion in a data set equates to full validation of that case, few subsequent users revisit a case's source materials to carefully consider the validity of any particular data point. Explicitly coding for such measures draws attention to the provenance of the data and makes this meta-data accessible to the analyst.
3. Explicitly coding for these measures allows for the inclusion in the data set, during the initial collection process, of borderline cases that can easily be excluded—if desired—by researchers during subsequent analysis.
4. In the absence of explicit coding, even if the analyst trusts that the database creators dutifully evaluated sources, there is no record that this "hidden" evaluation process was applied consistently across all cases. Explicit measures create confidence that claimed evaluation procedures were indeed followed by coders.
5. The process of validation and the explicit inclusion of metrics will likely encourage the use of a broader range of sources (rather than relying on a single source), which usually results in a more robust data set.

## *Source Evaluation in Practice: Open-Source Event Data Sets*

Despite these benefits, in at least one social science area—International Relations—the authors found scant use of any kind of explicit measure, and only sporadic mention of any type of source evaluation (see Table 1 for an illustrative sample of data sets). To begin with, several of the major international event data sets rely on a single source or may have started with adequate comparative sources and reduced source scope, which, as discussed above, is strongly recommended against. Importantly, by doing so, such data sets inherently adopt the same selection bias and objectivity level as the source upon which they rely, potentially

**Table 1.** Source evaluation in existing data sets.

| Data set | How source evaluation issue is addressed |
|---|---|
| 10 Million International Dyadic Events | Relies solely on *Reuters*[56] |
| World Event Interaction Survey (WEIS) | Relies solely on the *New York Times*[57] |
| American Terrorism Study, 1980-2002 | Based on court records, followed up by systematic media searches[58] |
| NCTC Worldwide Incidents Tracking System (WITS) | Website states that the cases are based on information "from a variety of open sources that may be of varying credibility"; sources are not provided publicly[59] |
| RAND Worldwide Terrorism Incident Knowledge Database | Actual database does not list source validity; all sources are from open source news agencies and two sources are generally required before a case is added; some incidents are apparently checked against classified sources[60] |
| Conflict and Peace Data Bank (COPDAB) | Lists 70 sources without rating credibility/validity[61] |
| Correlates of War (COW) Militarized Interstate Disputes (v3.10) | Lists major sources and lists source for each case without rating credibility/validity[62] |
| Uppsala Conflict Data Program (UCDP) Battle-Deaths Data set | Internally evaluates the independence and transparency of open sources based on context and attempts to trace back to primary sources; data from biased and unreliable sources is averaged only into their "high" measure; does not provide an explicit measure[63] |
| International Terrorism: Attributes of Terrorist Events, 1968-1977 (ITERATE 2) | Available data set does not explicate source validity[64] |
| Global Terrorism Database (GTD) | Uses an (unpublished) four-point scale to evaluate source validity, but measures are not attached to publicly available data[65] |
| United States Extremist Crime Database (ECDB) | Based on existing databases, official government sources, academic chronologies, watch-group publications, and media sources. More "trusted" source(s) receives more weight when sources conflict[66] |

reducing overall confidence in the validity of the data and requiring any analysis to include appropriate caveats.

Even though some international relations data sets might use multiple sources and list the major sources of the data used for each case, few assess the quality of their sources in any way and almost none do so transparently and/or explicitly, instead simply asserting that they use credible sources. For example, the RAND Worldwide Terrorism Incident Knowledge Database generally requires at least two open sources per case, but does not provide an explanation of source credibility or case validity beyond surface corroboration. The Uppsala Conflict Data Program Battle-Deaths data set claims its sources are evaluated but does not present a coded variable for use by researchers. The United States ECDB settles "discrepancies" by giving more weight to the more "trusted" source; for transparency, the ECDB provides a ranking of source type reliability that determines relative trustworthiness. However, these decisions are not explicitly coded for each crime. Finally, the GTD employs a basic scale to evaluate the information used; however, this scale is used for internal purposes only and is not published with the rest of the dataset, which makes it inaccessible to the general researcher. Moreover, we suspect that because—based on correspondence with GTD managers—this evaluation measure appears to not differentiate conceptually or practically between the component elements of validity and credibility as we have defined them, the measure is likely to be of only limited utility. But the lack of an explicit measure in the publicly available data precludes testing of this hypothesis. In sum, no current conflict event data set of which the authors are aware—let alone those observing terrorism—provides a comprehensive means for explicitly rating source credibility, data validity and overall case validity for use by researchers. This meant that existing event databases proved to be of only marginal utility in guiding the authors toward formulating a comprehensive and systematic approach to evaluating source credibility.

## Toward A Transparent Approach to Source Evaluation

The above discussion has introduced a number of aspects of source evaluation and evinced a need for following transparent procedures in this regard. To allow for the greatest flexibility in subsequent analysis, we have argued that it is desirable to explicitly code for the credibility of sources, and thence for the validity of the overall event, as well as specific variables. Once this general strategy had been laid out, the next task we faced was to operationalize the various source evaluation measures in terms of codable metrics. Many of the same issues arose as those that researchers confront when crafting any new data set, including: Which elements should be coded directly (as "objective" measures) and which should be evaluated or combined contextually by coders (as "subjective" measures)? How to balance specificity versus burdening the data set with too many additional variables? To what degree is quantification necessary *contra* the alternative of capturing qualitative descriptions? Which aspects warrant closer treatment and which can be subsumed within broader measures?

To answer these questions, we established a reference point by returning to first principles and reviewing exactly what we hoped that the measures would achieve in terms of addressing our concerns about the credibility and validity of the open source data. We therefore centered our efforts around the following framing question: *How much confidence do we have that the event genuinely occurred as described?* Parsing this question provided initial guidance as to the scope and nature of the required metrics. *How much* suggested at least some degree of quantification would be necessary, while the desire for *confidence* implied some type of scale to measure source credibility and data validity. The need to identify a *we* (the person whose perspective is being considered) drew our attention to the envisaged users of the data set and emphasized that the measures must be not only robust, but user-friendly and easy to understand. Since we sought the derived validity of open sources with reference to a particular *event*, our measures would also have to contend with the precise nature of what an event entailed. Importantly, our focus was on what *genuinely occurred*, which indicated that our basis for evaluation should be ontological proximity to the ground truth. Last, the phrase *as described* cautioned that not only did we need to consider whether the event actually occurred (overall case validity), but also to some degree the details of the event. The framing question further implied that in addition to issues of source credibility, our measures should also, if possible, help address the problems of information insufficiency and inherent uncertainty that often plague event data collection from open sources.

Drawing on this exercise, as well as the theoretical aspects of credibility and validity described above, we crafted a Source Evaluation Schema, consisting of a set of operationalized variables and instructions. Several of these variables focus on capturing the intentional and accidental distortion of information. By recording these credibility-related variables for each open source used and including corroboration information, it is possible to arrive at a combined source-derived validity metric for the event, which can then be utilized on its own, or in conjunction with the individual source credibility metrics. An additional variable was included to reflect inherent uncertainty about the event, as described above in the section on the need for source evaluation. The last portion of the Schema involved assigning two basic validity metrics[31] to many of the substantive event variables to reflect uncertainty regarding the details of the event. It should be borne in mind that most of the measures we use are at least partly subjective in nature, in that they reflect the coders' beliefs about the information they have collected. Yet, we endeavored to make them as objective as possible and at least no more subjective than the codes for most of the substantive variables. The Source Evaluation Schema, including examples, is reproduced in its entirety in Appendix A, and we discuss the various components below.

### *Individual Source Credibility*

The variables in this component of the Schema are coded for each source (in our case, mainly printed news stories, but sometimes also court documents or government reports) in every case/event.[32] Coders employ intrinsic, as well as extrinsic, evaluation to code both of the following variables. Any salient bias or incompetence shown by either the author or publishing institution is regarded as tainting the credibility of the source as a whole. Authors and institutions are not evaluated separately since prejudices or deficits from either provide justification for skepticism and thus a lower rating. The first credibility metric is *Institutional and Author Objectivity*.[33] Rating the objectivity of a source provides a measure of the extent to which the provided information reflects bias. If either one of the author or the institutional publisher is biased, the source is regarded as biased. Owing to the coder subjectivity involved, a relatively broad (3-point) scale is used. The second credibility metric is *Institutional and Author Competence*,[34] which assesses the capability for accurate recording and reporting of information that an author/publisher brings to the event subject.

There is no single authoritative source that analyzes the quality, objectivity, and competence level of news sources. Therefore, to assist coders in extrinsically evaluating the general reputation of sources, we set up a procedure whereby experienced researchers would be consulted,[35] including (a) regional experts, who could be asked about the level of bias and competence in international sources and (b) researchers working on other data sets, who could give advice on source reputation. When consultants could not provide extrinsic evaluation advice, the coder was instructed to evaluate the publisher based on research of the publication record. For Web-based sources, certain potentially more reliable domains (e.g., .edu and .gov) often provide initial knowledge about the source.[36] In any event, the general reputation of sources not previously encountered by consulted researchers was built on cumulatively as each particular source was encountered repeatedly and an intrinsic evaluation was carried out.[37] The project therefore maintained and continually added to and modified a list of institutional publication sources and associated presumptive extrinsic credibility measures.

### *Overall Event Validity Determination*

This metric seeks to provide a cumulative measure of source-derived validity and—through a simple yet objective process of corroboration—connote the confidence we have that the event described by the sources actually occurred. The variable is coded once for each case and the key measure is the number of independent sources that record the event.[38] It is also a far more objective measure than the individual source-derived validity measures.

Crucially, if *either* the objectivity *or* competence metric of any source considered is coded as low, that source is not included in the calculation of the overall source-derived validity variable. This reflects our belief that where a source is predominantly biased or incompetent, it provides no substantive corroborating evidence and thus does not contribute towards an event's overall source-derived validity. Each source credibility measure is thus incorporated implicitly into the overall credibility measure. It should be noted that two sources displaying the same bias are not regarded as independent for the sake of the source-derived validity measure. Additionally, multiple sources from the same institution or author are only considered to constitute a single independent source.[39]

### *Inherent Event Uncertainty*

Some events are clouded by inherent uncertainty (i.e., unrelated to the credibility or number of sources). These usually reflect cases where the reporting is undisputed, but the larger question remains with respect to whether the event that occurred actually qualified for inclusion in the data set. An example drawn from our data set is a case where all sources (and the authorities) agree that there were dozens of deaths from a specific disease, but there was insufficient evidence to determine with any certainty whether the event represented a natural outbreak or an intentional attack. Since this variable pertains only to whether the nature of the event qualifies it for inclusion in the data set and not uncertainties regarding specific details of the event, it is coded once for each case.

### *Event Detail Evaluation*

We adjoined two dummy variables, representing the lack of corroboration and discrepancy, respectively, to most of the substantive data set variables, whether the original variable was qualitative or quantitative. So, for example, for each of our location variables we created two new variables (Country-Uncorroborated, Country-Discrepancy, City-Uncorroborated, City-Discrepancy, etc.). While important in our research, this level of detail (and hence effort) in source evaluation may not be necessary in all circumstances, and therefore its application to any particular variable is optional.

Providing even relatively basic variable-by-variable measures that indicate the degree of source support for the variable values allows the users of the data set the option of either including or excluding those cases on which sources differ as to the value of the variable, or where there is no corroborating source for a particular variable value. This imbues analysis with tremendous flexibility and enables extensive sensitivity analyses to be performed according to the level of confidence in the validity of the values listed for any particular variable.

## Illustrative Application

As a demonstration of the Source Evaluation Schema, we present some illustrative results from the application of the Schema to our event database. The POICN Database collects open source data on all politically or ideologically motivated incidents (from plots to actual uses) involving CBRN weapons from 1990 to the present. The database codes for 139 core variables relating to geospatial, temporal, motivational, tactical, and consequence aspects of each incident. In addition, following the above schema, the research team rated each source connected with an incident for credibility and each case for overall validity, together with providing a measure for the validity of 84 of the core variables.

For the 449 incidents initially coded (spanning all relevant events in the period 1990–2010), Figures 1a and 1b depict the relative frequencies of the source-derived validity and event uncertainty measures. Table 2 provides a summary of the credibility measures across all collected sources and Table 3 provides a frequency distribution of the detailed variable-level validity measures for four sample variables.

With the exception of the discrepancy metric, most of the evaluation measures display considerable variance across cases, and provide an initial indication that the Schema is indeed measuring a real phenomenon. It is also apparent that the separate measures of objectivity and competence provide unique information. The variation in objectivity and competence scores further illustrates the need for transparent evaluation. The reason for this is because, while High Objectivity and Full Competence sources are preferred, lower ranking sources often present additional information that may be useful for contextual or qualitative purposes, as long as such cases are treated skeptically during aggregate analyses. Connecting data directly to transparent source credibility ratings allows the data to remain valid for quantitative analysis while still including many cases of potential informative value.

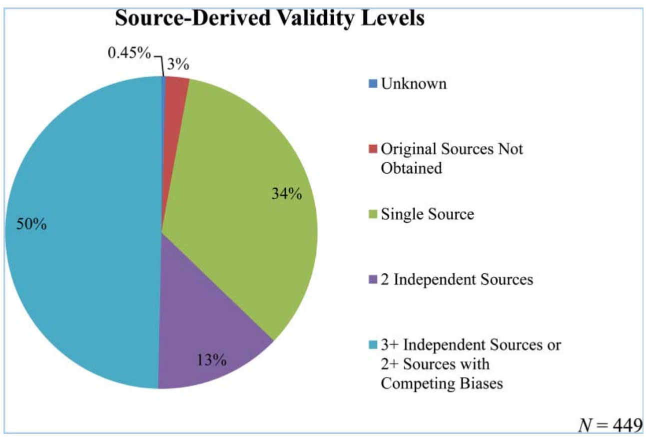


Figure 1a. POICN database source derived validity.

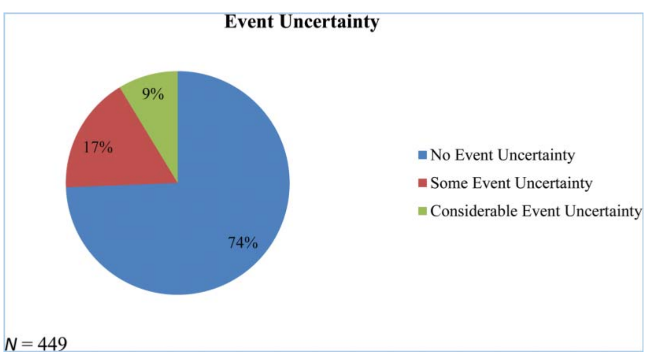


Figure 1b. POICN database event uncertainty.

**Table 2.** POICN database source credibility measures.

| Objectivity level of sources | | Competence level of sources | |
|---|---|---|---|
| **High** | 1253 (38.1%) | **Full** | 1,223 (37.2%) |
| **Potential** | 612 (18.6%) | **General** | 599 (18.2%) |
| **Low** | 38 (1.2%) | **Questionable** | 79 (2.4%) |
| **Inherited** | 781 (23.7%) | **Low** | 2 (<0.1%) |
| **Not Obtained** | 606 (18.4%) | **Inherited** | 780 (23.75%) |
| | | **Not obtained** | 607 (18.4%) |
| **Total** | 3290 (100%) | | 3290 (100%) |

Currently, very few Low Competence sources have been included in the database. However, as the database continues to evolve and an even broader range of source types is included, this rating may become particularly applicable for less traditional data sources, such as those presented through social media or blogs.

As to the salience of these measures, even from a purely descriptive comparison, it is immediately apparent that, for certain variables at least, case validity can have a substantial impact on visible outcomes. For instance, when considering cases with all levels of source-derived validity compared to only cases with high source-derived validity, a 13 percentage point change in apprehension rate[40] is observed (Figure 2). Additionally, Figure 3 shows the descriptive relationship between region and event type. The shaded portion in each bar represents the high source-derived validity cases, and the unshaded portion the lower-level validity cases. When observing all levels of source-derived validity versus only high source-derived validity, there are many shifts in which regions and event types are most relevant. For instance, in Russia and Eastern Europe, (Possession of a) Weapon more than doubles when the lower source-derived validity levels are included. These observable shifts at the very least suggest that including only cases with specific validity levels may significantly alter the outcomes of data analysis.

We tested the potential impact of these measures on statistical inference by running preliminary models with the data, while varying one of the evaluation measures, source-derived validity. Table 4 shows the results of running a series of simple regression models on all 449 cases in the database.[41] These models are put forward solely for the purpose of illustrating the use of the Source Evaluation Schema as opposed to substantive analysis and, as such, should not be used to guide CBRN or counterterrorism policy.[42] The models assess the effects of the region in which an event occurred and the nature of the perpetrator on the level of CBRN event (ranging from plots to actual weapon use), using a basic ordinal regression. Without entering into detail regarding the substantive underpinnings of the models, it is theoretically plausible that CBRN events occurring in different regions might progress to more or less "serious" levels, or that actors espousing different ideologies might enjoy different levels of success in their pursuit of CBRN weapons.[43] Model 1, the base model, includes all

**Table 3.** Event detail evaluation (four sample variables).

| Variable | Uncorroborated | | Discrepancy | |
|---|---|---|---|---|
| | Yes | No | Yes | No |
| **Event type** | 152 | 297 | 19 | 421 |
| **Heightened interest** | 146 | 303 | 3 | 446 |
| **Delivery** | 156 | 293 | 8 | 441 |
| **Attack date** | 83 | 366 | 5 | 444 |

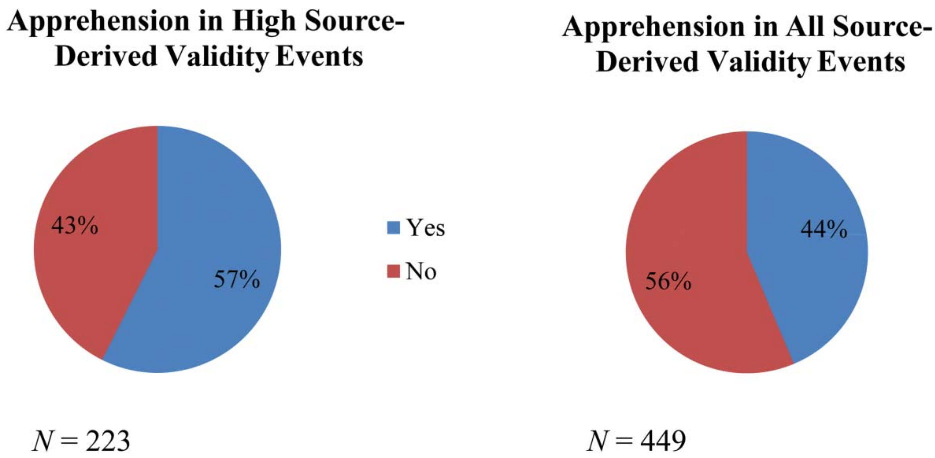


**Figure 2.** Apprehension rates across source-derived validity.

cases,[44] whereas Model 2 includes only those cases with the highest level of source-derived validity (i.e., a "3" on our scale). Model 3 is similar to Model 1, but includes source-derived validity as an explicit factor in the analysis, as well as an interaction term.

First, when all cases are considered, irrespective of their source-derived validity scores, all the types of perpetrator ideology are significant at the 5 percent level, and appear to be associated with lower level incidents (negative coefficient). However, in Model 2, with only cases receiving the highest source-derived validity rating included, Cults and Left-Wing actors are no longer significant. At the same time, whereas in Model 1 Asia was the only region that was significantly (and positively) associated with the type of event, when only the highest

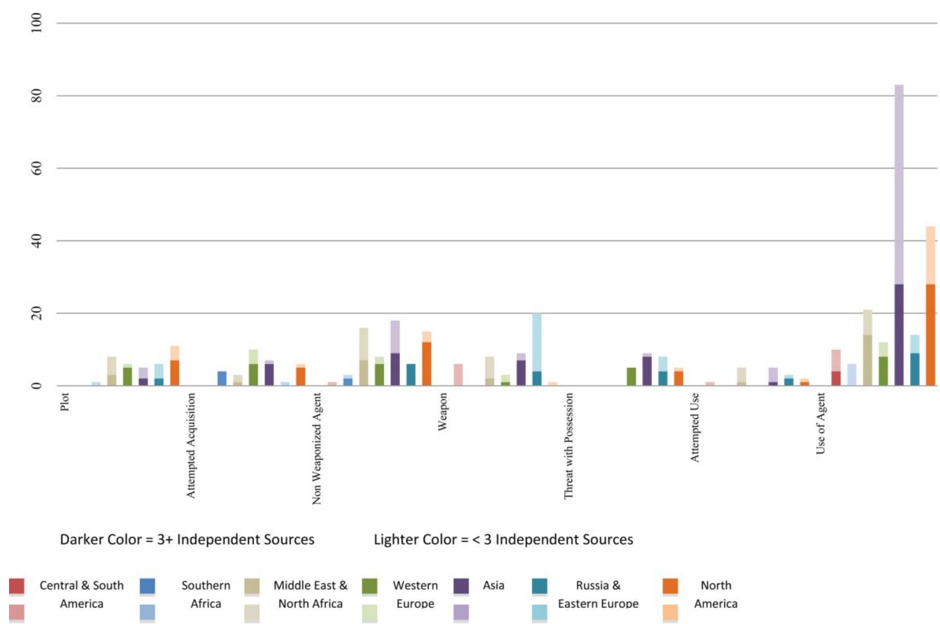


**Figure 3.** Source-derived validity by region and event type.

**Table 4.** Factors influencing level of event.

| Results of Ordinal Regression (OLOGIT) | | | | | | |
|---|---|---|---|---|---|---|
| | Model 1: All cases ($n = 449$) | | Model 2: High validity ($n = 223$) | | Model 3: All cases plus validity and interaction terms ($n = 436$) | |
| Variable | Coefficient (SE) | Significance ($p$ value) | Coefficient (SE) | Significance ($p$ value) | Coefficient (SE) | Significance ($p$ value) |
| Lone Actor | −2.266 (.350) | **0.000** | −1.772 (.447) | **0.000** | −2.428 (.386) | **0.000** |
| Cult | −1.934 (.445) | **0.000** | −1.608 (.604) | **0.008** | −2.171 (.460) | **0.000** |
| Religious Extremist | −3.363 (.361) | **0.000** | −3.178 (.499) | **0.000** | −3.778 (.385) | **0.000** |
| Right Wing | −3.471 (.593) | **0.000** | −2.497 (.922) | **0.007** | −3.710 (.607) | **0.000** |
| Left Wing | −1.144 (.496) | **0.021** | −1.311 (1.31) | 0.317 | −1.177 (.501) | **0.019** |
| Ethno-Nationalist | −1.226 (.366) | **0.001** | −1.870 (.621) | **0.003** | −1.333 (.380) | **0.000** |
| South/Central America | 0.489 (.586) | 0.404 | 16.270 () | | 0.416 (.590) | 0480 |
| Sub-Saharan Africa | 0.074 (.553) | 0.894 | −1.103 (.765) | 0.149 | 0.088 (.561) | 0.875 |
| Middle East/North Africa (MENA) | 0.810 (.363) | 0.026 | 2.135 (.555) | **0.000** | 0.052 (.425) | 0.902 |
| Western Europe | 0.028 (.343) | 0.935 | 0.298 (.425) | 0.484 | 0.051 (.352) | 0.884 |
| Asia | 1.238 (.312) | **0.000** | 1.121 (.405) | **0.006** | 1.280 (.320) | **0.000** |
| Eastern Europe/ Russia | −0.328 (.384) | 0.393 | .773 (.602) | 0.199 | −0.439 (.394) | 0.265 |
| High Source-derived Validity (SDV = 3) | ~ | ~ | ~ | ~ | 0.091 (.220) | 0. 678 |
| SDV × MENA | ~ | ~ | ~ | ~ | 2.130 (.545) | **0.000** |

Note: Dependent variable does not vary over any cases where independent dummy = 1, resulting in zero standard error and undefined significance level. Ordinal dependent variable: Plot; Attempted Acquisition; Possession of Agent; Possession of Weapon; Threat with Possession; Attempted Use; Use of Agent. $p$-value for two-tailed test; ~ = not applicable.

credibility cases are included in Model 2, the Middle East/North Africa region is now highly significant (<0.1 percent) with a moderately large positive coefficient. The obvious response to these models is that, since the cases in Model 2 constitute a subset of Model 1, the regression is being run on different samples of different sizes and therefore that observed changes in the significance of variables do not necessarily imply an improvement in the model. This is essentially correct—and we will turn to the question of which model is "better" below—but one must not forget that the very fact that the two models differ as to which independent variables are salient relative to the outcome is non-trivial. This is because—in the absence of a coded validity measure—from the point of view of a user of the database, only one of these regressions would ever be run. In the absence of any source evaluation, the regression would yield the results of Model 1, while, even in the case where the creator of the data set performed a systematic, but implicit source evaluation, results presumably similar to those in Model 2 would be obtained. In either event, the user would not "see" the other model, would remain unaware of any changes in the significance of variables, and thus not be in a position to even begin to look more closely at the affected variables.

To begin to understand why Models 1 and 2 yield different results, we need to examine the different subsamples more closely. From this point on we will focus our analysis on the Middle East/North Africa (MENA) variable, partly because there is a widely held belief in a significant geographic reporting bias in the open sources.[45] In Figure 3, little can be learned from the proportion of total MENA cases that receive a high validity score (44 percent), since this proportion is similar to that of several other regions (Sub-Saharan Africa, Russia and Eastern Europe and Asia). However, when one looks at the relative distribution of MENA cases across Event Types, it can be seen that the high-validity subsample has a greater proportion of MENA cases at the far end of the scale (Use of Agent) relative to the other regions. In other words, relatively fewer MENA Uses of Agent are discarded in moving from high validity to lower validity cases than, say, Asian Uses of Agent. This provides at least an initial explanation for the change in significance of the MENA variable.

To take this further and assess whether or not this is capturing a real effect as opposed to a spurious artifact of sampling, we turn to Model 3. The first thing to notice in this model is that the Source-Derived Validity score, when included as a variable in the model, is not significant. Although at first this might appear to undermine the relevance of the validity variables, upon closer reflection this result is unsurprising since there is no inherent reason for less-credible sources to more commonly report on one type of event over another in a monotonic fashion. For example, biased or incompetent sources may be more or less likely to report failed attempts at acquisition of CBRN weapons compared with threats following successful acquisitions.[46] This does not mean, however, that the validity measure is not relevant, because it might still have an effect through interactions with other variables. Since we are focusing attention for now on the MENA variable, we interact MENA with a dummy variable for high source-derived validity (in essence, highlighting in the regression those MENA cases with high validity scores) and include this term in the regression. We see from Model 3 that the interaction term is highly significant and positive, thus suggesting that, despite the weak association overall between source-derived validity and event level, those MENA cases with higher validity are more likely to be higher up the event level scale, all else being equal, than the reference region (North America). Ignoring source-derived validity completely in the analysis would thus likely result in an analyst missing the potentially substantial impact of the MENA region as a contributor to the threat. Together, the three models appear to provide various indications that including source evaluation considerations in the analysis can help to provide a more complete analysis.

If we assume that most reporting bias in the CBRN topic domain tends toward sensationalism and thereby the inclusion of non-genuine cases, we must accept that the possibility of false positives in the data increases. These examples demonstrate this effect and illustrate how the Source Evaluation Schema can help avoid drawing potentially spurious, or at least, incomplete, conclusions from the event data. In essence, the Schema addresses inherent reporting biases in event data by introducing a corrective selection bias.

## Broader Implementation and Conclusion

The illustrative application argues for the utility of the Schema within the authors' CBRN terrorism event data set; yet, an interdisciplinary conversation on the general effectiveness of such a schema is needed, especially with respect to expanding and clarifying the process of implementation. A schema of this kind can be implemented in both new and existing data sets in many fields reliant on open sources of varying objectivity and competence. It can also be applied to event data from a variety of time periods, geographic locations, and reflecting the actions of a variety of actors. In many— but not all—circumstances, the efficacy of clarity and reliability provided by schema-enabled data analysis will certainly justify the resource requirements of implementation. The benefits are most likely to be worth the implementation costs when dealing with challenging and controversial open source reporting, such as that focused on conflict, terrorism, or another politically charged arena. In all cases, the actual amount of resources required for implementation is proportional to the complexity and scale of the data set, along with the number and diversity of the sources used. As might be expected, coding for the additional variables of source-derived validity, event uncertainty and variable-level validity increases the time taken to construct the database in proportion to the number of events in the data set. Yet, the procedure of including source evaluation variables, though it diverges somewhat from the coding of other event characteristics, does not necessarily increase the coding burden more than adding traditional event-level variables. Incorporating source evaluation codes can thus be compared favorably with any other expansion of a data set and in many cases would add greater value to the overall data than adding more event descriptors, say, a more detailed location variable. Performing intrinsic and extrinsic evaluations of each source does, however, introduce genuinely new resource requirements in terms of time and expertise beyond that which would be used to simply add new event-level variables.

In the example of the POICN database, we found that while the coders at first found implementation of the Schema somewhat removed from their previous experience, they quickly became familiar with the procedures and after a little practice became adept at coding source credibility and validity measures for each case.[47] Over time, our training of coders has become much more streamlined, allowing for schema implementation with more efficiency. The source evaluation adds approximately 10 percent more time to case coding once a coder has been trained in source evaluation and has an understanding of the relevant regional source institutions. The amount of time needed to code discrepancy and lack of corroboration for each variable depends on the complexity of conflicting sources.

The authors are not proposing that the Schema described in the Appendix should be adopted *in toto* by all collectors of open-source data; rather what we present can serve as a blueprint that can be modified within different disciplines and implemented to varying extents depending on needs. The Schema can also complement other approaches to increasing the validity of data usage, such as conducting analysis across multiple similar event datasets.[48]

The proposed schema could be tailored in several ways based on subject matter and resources. While resource and time constraints may limit the full schema implementation in some circumstances, even implementing select pieces of the Schema can increase evaluation transparency and analysis robustness. For instance, for terrorism-related events data, the authors would suggest utilizing the uncertainty measure. This requires few resources beyond the conceptualization of what uncertainty means in context and could prove useful in many disciplines. Further, coding validity at the variable level can be time intensive. A project with greater time constraints may choose to bolster only the most theoretically relevant variables with the corroboration and discrepancy coding.

Integrating the Schema into an already existing data set may require other accommodations. For instance, implementing the Schema within a data set that does not collect source citations would be resource intensive, akin to recoding the entire dataset. However, in a data set that has collected source citations but not the actual sources, it may be possible to create a credibility metric based on the relative objectivity and competence of the source institutions rather than the specific sources. Moreover, as sources for events data become more diverse, the Schema is sufficiently flexible so as to allow for the inclusion of new media.

Two collaborative steps would allow for efficient and effective implementation of tailored versions of the Source Evaluation Schema within any social science discipline that uses highly politicized event data gathered from open sources. First, as researchers in other disciplines adjust the Source Evaluation Schema presented here for discipline-specific issues, these new schemas could be posted on a central website. A library of easily accessible schemas would enable interdisciplinary critiques and facilitate the introduction and sharing of schemas across relevant communities. Second, and perhaps most importantly, we advocate for the creation of a Source Evaluation Repository. An international group of media researchers and others who study the proclivities of media institutions and other publishing entities would create a forum to crowdsource a rating baseline for commonly used publications. While such ratings may be somewhat subjective, the repository would provide coders and researchers with initial departure points for source ratings. In addition, following Strawn's advice,[49] the repository could provide coverage tendencies relative to other source institutions. Over time, additional users would increase the reliability of the ratings and historical ratings may show trends in the credibility of some source institutions. Such a repository should also include ratings for evolving media sources such as blogs and tweets. Moreover, the metrics should account for the use of international media without showing a bias for Western sources.

In the future, additional source evaluation strategies may be needed for a wider range of sources, including blogs and personal websites. No matter which aspects of the Schema are implemented, objective and transparent metrics will provide greater flexibility, transparency and therefore value to researchers.

After establishing the disconcerting lack of explicit measures in extant open-source event data sets, we have put forward a schema for evaluating the credibility of sources and their effect on overall event validity. Although we have tried to present a practicable set of measures, we do not assert that our Schema is necessarily the best implementation of source evaluation and indeed encourage further development of these ideas. Based on our preliminary comparisons using these measures in a real-world application, we do, however, argue that the inclusion of such measures in similar datasets can facilitate more robust and defensible analyses and thus ultimately strengthen the role that social science can play in guiding and improving policy choices, especially in highly charged political and security contexts like terrorism.



## Acknowledgments

The authors thank John Sawyer for his assistance in creating the POICN database and implementing the Source Evaluation Schema presented here. They also thank Phillip Schrodt, Erin Miller, Laura Dugan and Amy Pate, together with several anonymous reviewers, for their helpful comments. Any errors or omissions, however, are solely the responsibility of the authors.

## Funding

This work was supported by the Defense Threat Reduction Agency, Basic Research Award #HDTRA1-10-1-0017. Any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the U.S. Department of Defense.

## Notes

1. This Voltaire quotation is from the *Letter to Le Comte d'Argental*, August 28, 1760, quoted in Laura Moncur's *Motivational Quotations*. http://www.quotationspage.com (accessed February 20, 2016). The quotation from Joe Moore was found on Quotes.net, STANDS4 LLC, 2016. "Joe Moore Quotes." http://www.quotes.net/quote/15830 (accessed February 20, 2016).
2. Stevyn D. Gibson,"In the Eye of the Perfect Storm: Re-Imagining, Reforming and Refocusing Intelligence for Risk, Globalisation and Changing Societal Expectation," *Risk Management* 7(4) (2005), pp. 23–41.

3. See descriptions of several recent government initiatives and related discussions at the website of The Coalition for Evidence-Based Policy. "Coalition for Evidence-Based Policy" (2011). Available at http://coalition4evidence.org/wordpress/ (accessed 27 January 2011).
4. We are not discussing the reciprocal problem, that is, the exclusion of genuine cases (false negatives or Type II errors), here.
5. As an example of the complexities of data validity in these highly charged issues, see Schmid for difficulties of evaluating information on terrorism. Alex P. Schmid, "Terrorism—The Definitional Problem," *Case West Reserve Journal of International Law* 36(2–3) (2004), pp. 375–419.
6. David Snyder and William R. Kelly, "Conflict Intensity, Media Sensitivity and the Validity of Newspaper Data," *American Sociological Review* 42(1) (1977), pp. 105–123.
7. Pamela E. Oliver and Gregory M. Maney, "Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions," *American Journal of Sociology* 106 (2) (2000), p. 495.
8. Andrew Reeves, Stephen Shellman, and Brandon Stewart, "Fair & Balanced or Fit to Print? Effects of Media Sources on Statistical Inferences." Paper presented at the annual meeting for the International Studies Association, San Diego, California, 22–25 March 2006.
9. Steven M. Chermak, Joshua D. Freilich, William S. Parkin, and James P. Lynch, "American Terrorism and Extremist Crime Data Sources and Selectivity Bias: An Investigation Focusing on Homicide Events Committed by Far-Right Extremists," *Journal of Quantitative Criminology* 28 (2012), p. 214.
10. In addition, selectivity checks and robustness of findings across data sets may help solidify conclusions. Enders et al. compare the events within the GTD and the International Terrorism: Attributes of Terrorist Events (ITERATE) data set to provide a method to calibrate GTD terrorist incidents (in which coding procedures shifted over time) to the historically consistently coded ITERATE data set. Walter Enders, Todd Sandler, and Khusrav Gaibulloev, "Domestic versus Transnational Terrorism: Data, Decomposition, and Dynamics," *Journal of Peace Research* 48 (May 2011), pp. 319–337.
11. Simon Hug, "Selection Bias in Comparative Research: The Case of Incomplete Data Sets," *Political Analysis* 11(3) (2003), pp. 255–274.
12. Philip A. Schrodt, "Precedents, Progress, and Prospects in Political Event Data," *International Interactions* 38(4) (2012), pp. 546–569.
13. Gary LaFree and Laura Dugan, "Introducing the Global Terrorism Database," *Terrorism and Political Violence* 19(2) (2007), pp. 188. For instance, Sandler's (1995) analysis of terrorism and democracy suggests a coverage bias: the lack of free press in autocracies results in less reporting on potential terrorist events in autocracies compared to democracies. Todd Sandler, "On the Relationship between Democracy and Terrorism." *Terrorism and Political Violence* 7(4) (1995), pp. 1–9.
14. Jennifer Earl, Andrew Martin, John D. McCarthy, and Sarah A. Soule, "The Use of Newspaper Data in the Study of Collective Action," *Annual Review of Sociology* 30 (2004), p. 72.
15. Christian Davenport and Marika Litras, "Rashomon and Repression: A Multi-Source Analysis of Contentious Events." (2003) *Working Paper*.
16. John T. Woolley, "Using Media-Based Data in Studies of Politics," *American Journal of Political Science* 44(1) (2000), pp. 156–173.
17. José Barranco and Dominique Wisler, "Validity and Systematicity of Newspaper Data in Event Analysis," *European Sociological Review* 15(3) (1999), p. 319.
18. Robert M. Entman, "Framing Bias: Media in the Distribution of Power," *Journal of Communication* 57 (2007), pp. 163–173.
19. Kelley Strawn, "Validity and Media-Derived Protest Events Data: Examining Relative Coverage Tendencies in Mexican News Media," *Mobilization* 13(2) (2008), pp. 147–164.
20. See, for example, Don Munton, "Policy Makers, Public Records and Reality," in D. Munton, ed., *Measuring International Behavior: Public Sources, Events and Validity* (Halifax, Canada: Centre for Foreign Policy Studies, Dalhousie University, 1978), pp. 67–92.
21. Russell Leng, "Events Data Validity: Comparing Coding Schemes," in D. Munton, ed., *Measuring International Behavior: Public Sources, Events and Validity* (Halifax, Canada: Centre for Foreign Policy Studies, Dalhousie University, 1978), pp. 127–141.

22. "Credibility," in *Merriam-Webster's Online Dictionary*. 11th ed. (2011). Available at http://www.merriam-webster.com/dictionary/credibility (accessed 12 January 2011).
23. William P. Cassidy, "Online News Credibility: An Examination of the Perceptions of Newspaper Journalists," *Journal of Computer-Mediated Communication* 12(2) (2007), pp. 478–498.
24. Sophia Peterson, "News Selection and Source Validity," in D. Munton, ed., *Measuring International Behavior: Public Sources, Events and Validity* (Halifax, Canada: Centre for Foreign Policy Studies, Dalhousie University, 1978), pp. 43–66.
25. For news media sources, a citation with the country, source, and date should be sufficient for replication purposes, with the ideal inclusion of the unique record identification number. Published books require a full citation and relevant page numbers. Records of Internet sources and social media should also include Portable Document Format (PDF) files or screen shots whenever possible.
26. See LaFree and Dugan for a description of coder training and data verification for the original GTD and Freilich et al. for an explanation of the coder training and interrater reliability used within the ECDB. Researchers overseeing the coding for both data sets use extensive training to help reduce the potential for coding errors. LaFree and Dugan, "Introducing the Global Terrorism Database," p. 186; Joshua D. Freilich, Steven M. Chermak, Roberta Belli, Jeff Gruenewald, and William S. Parkin, "Introducing the United States Extremist Crime Database (ECDB)," *Terrorism and Political Violence* 26(2) (2014), pp. 374–375.
27. This article does not attempt to address the issue of determining what events are never reported at all, an issue that requires qualitatively different remedial procedures. Non-reporting due to regional source limitations exacerbates selection bias; for example, some countries may receive limited database inclusion due to lacking an independent press.
28. Hug, "Selection Bias."
29. Elizabeth Kirk, "Evaluating Internet Information," *John Hopkins University* (2006). Available at http://www.library.jhu.edu/researchhelp/general/evaluating/index.html (accessed 29 June 2010); Joan Ormondroyd, Michael Engle, and Tony Cosgrave, "Critically Analyzing Information Sources" (2009). Available at http://www.library.cornell.edu/olinuris/ref/research/skill26.htm (ccessed 28 June 2010).
30. Soo Young Rieh and David R. Danielson, "Credibility: A Multidisciplinary Framework," in B. Cronin, ed., *Annual Review of Information Science and Technology*, Vol. 41 (Medford, NJ: Information Today, 2007), pp. 307–364.
31. These two metrics assess the validity of each variable indirectly, by assessing the number of sources that corroborate the data for each variable.
32. Freilich et al. and Sageman discuss ranking the relative reliability of source types. However, the Schema transparently codes each source allowing for varied objectivity and competence even within source type. Freilich et al., "Introducing the United States Extremist Crime Database"; Marc Sageman, *Understanding Terror Networks* (Philadelphia, PA: University of Pennsylvania Press, 2004).
33. The objectivity of each source is rated as High, Potential, Low, Original Source Unavailable, or Inherited.
34. The competence of each source is rated as Full, General, Questionable, Low, Original Source Unavailable, or Inherited.
35. While space limitations preclude a full description of this process, see Ayyub and Rowe and Wright for discussions on eliciting and aggregating the estimates of experts. Bilal M. Ayyub, *Elicitation of Expert Opinions for Uncertainty and Risks* (Boca Raton, FL: CRC Press, 2001); Gene Rowe and George Wright. "Expert Opinions in Forecasting: The Role of the Delphi Technique," in J. Scott Armstrong, ed., *Principles of Forecasting* (New York: Springer-Verlag, 2001).
36. Susan Beck, "The Good, The Bad & The Ugly: or, Why It's a Good Idea to Evaluate Web Sources," *Evaluation Criteria* (1997). Available at http://lib.nmsu.edu/instruction/evalcrit.html (accessed 30 June 2010).
37. Following Rieh and Danielson, "Credibility."
38. A source is regarded as independent of another source if it does not share the same original authorship and does not rely on the same original secondary source material. Additionally, if a coder uncovers multiple sources from the same institution (such as the *New York Times* or

Associated Press), only the most recent source is counted as an independent source, with the earlier sources coded as "Inherited" and linked to the most recent source.

39. Admittedly, in many cases determining the original author may not be possible. For instance, when international sources use local stringers to produce content, coders may not be able to determine when a stringer originally produced a story or when a single stringer produced similar stories for several news outlets.
40. Apprehension = 1 when at least one perpetrator was arrested.
41. Within the Schema, only two variables are coded at the event level: Source-Derived Validity and Inherent Event Uncertainty. In 449 cases, these two variables have a correlation of –.1179. When the analysis is run conditioned on Inherent Event Uncertainty (none) instead of Source-Derived Validity (high), the significance levels are very similar, except Left-Wing actors are statistically significant.
42. Given their illustrative nature, we present basic models and do not discuss model validation procedures (such as goodness of fit).
43. Nonetheless, if we choose not to make the proportional odds assumptions and instead non-parametrically use ordinary least squares to better understand the relationship between variables, we find that almost all variables remain significant at the same levels.
44. The reference categories in the models are Single Issue/Criminal Organizations and North America.
45. Similar analyses could also be conducted on the other two variables which had substantive changes in levels of significance.
46. Descriptively, there is no clearly evident relationship, with, for example, spikes in the third and seventh ordinal event categories, and a big dip in the sixth category. Although there is some significant two-way association between source-derived validity and level of CBRN event (Chi-square is significant to three decimal places), various measures of the magnitude of association (e.g., Gamma, Kendall's tau-b) are all less than absolute value 0.13, suggesting a very weak association overall between these variables. It is therefore little surprise that this validity measure is not a significant variable in the regression.
47. To assess how consistently coders could implement the Source Evaluation Schema, we supplied each of seven new coders, as an exercise during their coder training, with a sample of ten cases and asked them to code the various evaluation metrics. For nine (90 percent) of the cases, at least 6 of the 7 coders agreed on the source-derived validity level, the main evaluation metric that incorporates both the competence and the objectivity scores. This demonstrated that the Schema was practicable in the case of the POICN database and lent weight to its feasibility in other event databases.
48. See, for example, Enders et al., "Domestic versus Transnational Terrorism."
49. Strawn, "Validity and Media-Derived Protest Events Data."
50. We believe that even if they do not share any common publication lineage, two sources displaying the same potentially distorting filter, essentially share the same political or ideological "cause" and therefore might be susceptible to common incentives to either consciously or unconsciously misinform. This implies that these should not be given the status of independent sources for the purposes of determining source-derived validity.
51. The reasoning behind this is essentially the converse of that reflected in the previous footnote and asserts that when two reporting entities with tendencies to distort in opposing directions agree, this raises the level of credibility of the event itself.
52. Note that in this case only, we consider the subjects contained in the sources, such as experts interviewed, statements by officials, and so forth.
53. Rather obviously, in those instances where all observers agree that the event occurred in a way that would not warrant inclusion in the data set, the event is excluded.
54. A reasonably credible source is defined as a source with the values *source_comp* $\geq$ 2 AND *source_object*$\geq$1.
55. While not the primary concern of this article, the discovery of a discrepancy required additional coding rules for deciding which of the proffered values to code for that variable. Our coding rules in such situations are as follows: Based on our experience, credible reports published some time after an event are likely to yield more robust data for many variables such as the number of fatalities and the perpetrator, so we use the value from the most recently published, reasonably credible source (providing it is published in a unique time period). If all sources were published in the

same time period, the information supported by the majority of reasonably credible sources is used. If the discrepancy is between only two sources within the same time period, the information from the source with the highest combined credibility rating is used.
56. Gary King and Will Lowe, *10 Million International Dyadic Events*. IQSS Dataverse Network (2003).
57. Llewellyn D. Howell, "A Comparative Study of the WEIS and COPDAB Data Sets," *International Studies Quarterly* 27(2) (1983), pp. 149–159.
58. Brent L. Smith and Kelly R. Damphousse, *American Terrorism Study, 1980–2002* (Ann Arbor, MI: Inter-University Consortium for Political and Social Research, 2007).
59. National Counterterrorism Center, WITS NextGen. *Worldwide Incidents Tracking System* (2011). Available at https://wits.nctc.gov/FederalDiscoverWITS/index.do?N=0 (accessed 8 March 2011).
60. RAND Personal Correspondence (2010). RAND Worldwide Terrorism Incident Knowledge Database. 23 June 2010.
61. Edward E. Azar, "The Conflict and Peace Data Bank (COPDAB) Project," *Journal of Conflict Resolution* 24(1) (1980), pp. 143–152.
62. Faten Ghosn, Glenn Palmer, and Stuart A Bremer, "The MID3 Data Set, 1993—2001: Procedures, Coding Rules, and Description," *Conflict Management and Peace Science* 21(2) (2004), pp. 133–154.
63. Nils Petter Gleditsch, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand, "Armed Conflict 1946–2001: A New Dataset," *Journal of Peace Research* 39(5) (2002), pp. 615–637.
64. Edward F. Mickolus, International Terrorism: Attributes of Terrorist Events, 1968–1977 [ITERATE 2]. ICPSR07947-v1. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [producer and distributor], 1982. http://doi.org/10.3886/ICPSR07947.v1
65. START Personal Correspondence (2011). Global Terrorism Database. 30 June 2010.
66. Freilich et al., "Introducing the United States Extremist Crime Database."

## Appendix A: Source Evaluation Schema

### *Individual source credibility*

1. Institutional and Author Objectivity (*source_object*)
   Rating the objectivity of a source provides a subjective measure of the extent to which the provided information reflects bias. If either one of the author or the institutional publisher is biased, the source is regarded as biased. Owing to the coder subjectivity involved, a relatively broad (3-point) scale was used.
   **–99 = Inherited**
   If the source is not independent, objectivity is based on the measure attributed to the original source.
   **0 = Low**
   Author and/or institutional publisher have consistently demonstrated systematic bias (extrinsic evaluation) *or* the source document clearly reflects a lack of objectivity (intrinsic evaluation), signified by such characteristics as overly emotive writing. Examples include: a newspaper directly affiliated with a terrorist group; a reporter with a history of advocating for or against a particular group without any use of facts; or a passage overtly sanitizing or exaggerating certain violent behaviors.
   **1 = Potential**
   No intrinsic indications of bias, but author and/or institution have demonstrated bias in some cases, but not others (i.e., non-systematically). For example, a newspaper that is generally measured in its approach to reporting but is known on occasion to take a very pro-Israeli (or pro-Palestinian) stance on the Israeli–Palestinian issue, a media

outlet owned by a member of a royal family that espouses his family's views or a state-run news source reporting on continuing conflict between rebels and the government.

**2 = High**

Neither the author nor the institution has a reputation for systematic bias and there are no intrinsic indications of bias, that is, the document itself shows no overt or easily recognizable signs of bias. To code "High" without prior knowledge of the author/publisher, the coder must research the history and reputation of the author and institutional publisher.

2. <u>Institutional and Author Competence</u> (*source_comp*)

   **–99 = Inherited**

   If the source is not independent, objectivity is based on the original source.

   **0 = Low**

   Extrinsic evaluation reveals that the author and/or institutional publisher: (a) have had serious and widespread questions raised about their reporting skills or (b) obviously lacked the resources or skills to have adequately reported on the event. Alternatively, intrinsic analysis of the source document indicates: (c) substantive inconsistencies or errors. Examples might include a tabloid with no reputation for serious journalism or an original language document with multiple overt typographical errors, misspellings, or grammatical errors.

   **1 = Questionable**

   While at least one of the institutional publisher or author has failed to develop a general reputation for high quality output (extrinsic evaluation), the source document itself shows a prima facie level of competence (intrinsic evaluation). Conversely, the institutional publisher and author are generally regarded as competent reporters producing high quality output, but the source document itself shows inconsistencies and/or errors of a non-negligible number or nature. An example would be a seemingly well-written and researched source from an institution that has previously published unique stories that have been heavily disputed by other media sources and never substantiated.

   **2 = General**

   The author and institutional publisher are generally regarded as producing high quality output (extrinsic evaluation of reputation) and the source document itself shows a prima facie level of competence (intrinsic evaluation). However, neither the author nor institutional publisher covers the subject matter area (often referred to as a "beat" in journalism circles) or geographic region on a regular basis.

   **3 = Full**

   Both the author and institution have demonstrated prior competence with respect to the geographical and substantive domain on which they are reporting and there are no intrinsic indications of a lack of competence.

### *Overall event validity determination*

1. <u>Source-Derived Validity (*sdv*)</u>

   Sources with **either** *source_object* = 0 **or** *source_comp* = 0 are excluded.

   **–99 = Unknown**

   No usable sources describing the event could be located.

**–88 = Not Obtained**
Source(s) exist but have not been accessed (includes translation and legal issues).
**1 = Single**
A single source or multiple non-independent sources describe the event.
**2 = Two Independent Sources**
Two independent sources describe the event and agree on the broad nature of the event. A vital check at this stage is whether these two ostensibly independent sources both display potential bias (i.e., *source_object* = 1). In this case the sources are revisited and each source's bias is compared. If both sources display the same bias, the sources are not regarded as independent for the purposes of source-derived validity.[50]
**3 = Three Plus Independent; Two With Competing Bias**
Three or more independent sources describe the event and agree on the broad nature of the event; **or** two independent sources with competing biases describe the event and agree on the broad nature of the event.[51] If two or more of the sources display any level of common bias, they are dealt with similarly to the above instruction.

## *Inherent event uncertainty*

1. Inherent Event Uncertainty (*uncert_event*)
   **0 = None**
   All sources agree that the event occurred in such a way that it warrants inclusion.
   **1 = Some**
   Most observers[52] believe that the event occurred in such a way that it warrants inclusion, but the sources portray some uncertainty regarding this.
   **2 = Considerable**
   The event most likely occurred in a way that would not warrant inclusion in the data set (in our example, either an accident or the result of natural causes), but the possibility has been raised by one or more observers that it may have constituted a genuine inclusion event (in our example, an attack).[53]

## *Event detail evaluation*

1. Detail Uncorroborated (*X_Uncorroborated*)
   **0 = None**
   More than one reasonably credible source[54] provides information on the detail variable.
   **1 = Uncorroborated**
   There is only a single reasonably credible source that provides a value for that detail variable.
2. Detail Discrepancy (*X_discr*)
   **0 = None**
   All reasonably credible sources provide the same information on the detail variable.
   **1 = Some**
   Two (or more) reasonably credible sources have a discrepancy in the information provided for a particular variable.[55]