Best Practices in Event Data Coding:
Improving Coding Quality in the Electoral Contention and Violence (ECAV) Data

Elio Amicarelli
Ursula Daxecker
University of Amsterdam

Introduction

The Electoral Contention and Violence (ECAV) project collects event data on election-related contention for all countries with competitive elections for the 1990-2012 period (Daxecker, Amicarelli, and Jung 2019). The disaggregation turn in conflict research has increased the attractiveness of event data because researchers can use them at the level of analysis most appropriate for their research design. While reporting bias is frequently recognized as a challenge in coding events from news reports (Hoglund and Oberg 2011; Chojnacki et al. 2012; Urdal 2008; Weidmann 2016), less attention is usually paid to coding reliability and validity. Yet as Ruggeri, Gizelis, and Dorrussen (2011) demonstrate, even if events are reported, and are reported accurately, important concerns remain regarding the quality of event data. The coding of events from news consists of two important steps; first that coders identify the same events from sources, and second that they interpret events similarly. For a project like ECAV, coders' ability to identify and encode events as outlined in the coding scheme is thus crucial to produce high-quality data. This research note aims to provide more detail and hence transparency about the process of event identification and encoding than other event data coding projects.

In this paper, we first discuss the objectives and strategy to assess event identification and encoding in the ECAV project. We then present a detailed assessment of event identification, followed by an assessment of coding reliability and validity. We conclude with recommendations on the usefulness of such assessments for coding scheme improvements and the training and selection of coders.

Objectives and Strategy

ECAV extracts large numbers of events from a very large number of news articles. ECAV includes more than 18,000 events extracted from almost 220,000 articles. Unlike other event data projects using machine coding (e.g. KEDS, El Diablo), the project relies exclusively on human coders. The coding process for event data requires that coders perform two important

steps, namely (1) *identifying* events of interest and (2) *encoding* each event with the above-discussed variables following a set of pre-established coding rules. Human coders can vary in their ability to correctly identify and encode events.[1] An assessment of coder quality is thus of crucial importance to ensure that they perform similarly in event identification and encoding.

Soon after 11 coders were recruited for the project, we developed an operational strategy to assess coding quality. To assess event identification and encoding, we created a dataset consisting of more than 100 ECAV events for the 1991 elections in India. This dataset was then used as a benchmark to assess coders' performance. We refer to these data as the Identification Gold Standard (IGS, the benchmark used in Stage 1) and Coding Gold Standard (CGS, the benchmark used in Stage 2). Throughout this process, coders were not told that the procedure was a test of their coding quality, but rather were informed that we were implementing a new procedure for the cross-validation of existing data. The exercise was conducted after coders have received the initial coding training, but before they started actual coding. The assessment phase took approximately two weeks of full-time work per coder. Results were used to improve the coding procedure, to provide extra training to coders based on their performance, and to inform decisions on coder retention. We discuss the assessment of both stages in detail.

Event Identification

To translate news text into ECAV event data, human coders are assigned a set of news articles. Coders read these articles and identify events relevant for the project. In this process, coders may fail to identify relevant events (false negatives) or may mistakenly identify irrelevant events (false positives). To assess performance, all coders were provided with the same set of articles for the 1991 elections in India and were asked to identify relevant ECAV events.[2] The events identified by each coder were then matched with the events in the Identification Gold Standard (IGS).[3] A binary vector containing a set of True Positives (coded as 1s) and False Negatives (coded as 0s) was then derived for each coder. This binary information was then used to calculate the following identification performance measures for each coder.

[1] There can be several causes for variation in coder ability such as *a)* different levels in the understanding of what has to be considered an event of interest, *b)* how the coding rules should be applied in order to translate the news reporting in machine-readable variables and *c)* the effort each coder puts in doing his or her job. Each of these can jeopardize the quality of the data.

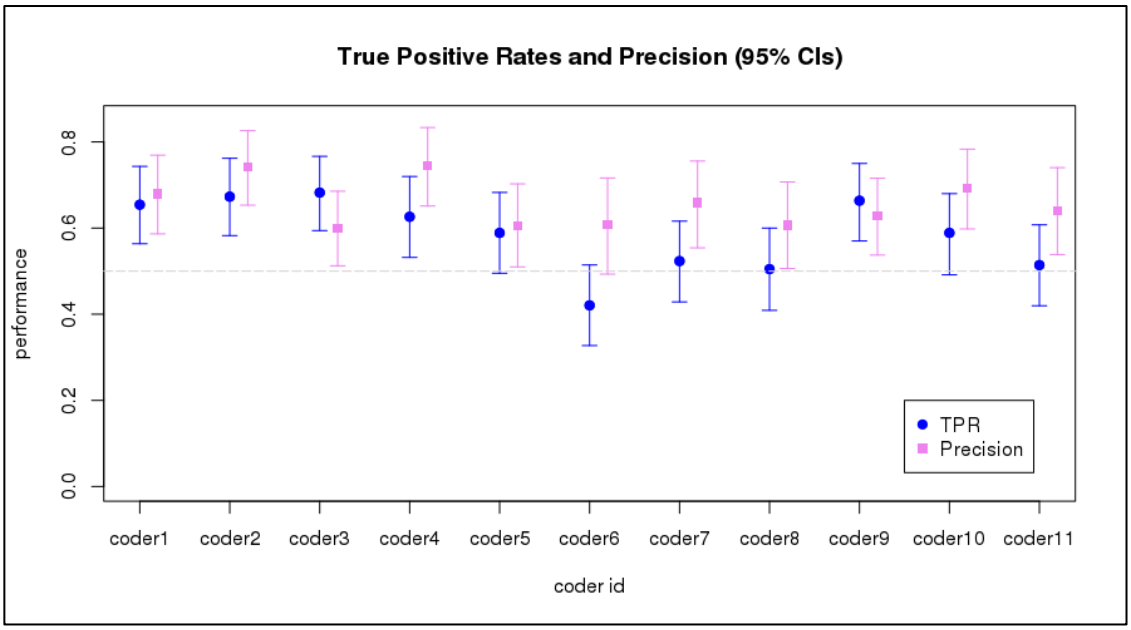
[2] For each identified ECAV event, coders use a spreadsheet where they record *a)* the relevant text snippet containing the event and *b)* the title, date and unique identifier of the article containing the text snippet.

[3] The IGS was based on hand coding by the PIs of the project.

- True Positive Rate (TPR): The TPR is the rate of correct identifications. The TPR ranges from 0 to 1. TPR equals 1 if a coder correctly identifies all ECAV events of interest; conversely, it equals 0 when a coder fails to identify any ECAV events of interest.[4]
- Precision (Prc) is the rate of identifications that turn out to be correct. Precision equals 1 if a coder is correct every time he or she identifies an event; conversely, it equals 0 when a coder is always wrong when identifying events. Unlike TPR, Precision is thus not sensitive to false negatives.[5]

The anonymized TPR and Precision results are showed in Figure 1. For both measures, 95% bootstrapped confidence intervals were calculated using the percentile method on 5000 bootstrap samples. The dashed grey line placed at a performance value of 0.5 represents the expected TPR performance of a balanced binary classifier producing random guesses.[6]

Figure 1: Individual True Positive Rate and Precision for 11 ECAV coders



As shown in Figure 1, coders 1, 2, 3, 4 and 9 have TPRs between 0.6 and 0.68 while coders 5,7,8,10 and 11 are between 0.5 and 0.59. The TPR for coder 6 is 0.42, the lowest in this set. A TPR of 0.42 means that coder 6 is correctly identifying 42% of the actual ECAV events

[4] $TPR = P(C_i = 1|IGS = 1) = TP/(TP + FN)$
[5] $Precision = P(IGS = 1|C_i = 1) = TP/(TP + FP)$
[6] We emphasize that the confidence intervals can be examined against the random theoretical expectation regardless of the degree of class imbalance only for the TPR (except for the unlikely case where the Gold Standard would only contain zeros), but not for Precision.

contained in the IGS.[7] Despite the fact that coders 1, 2, 3, 4 and 9 are those doing a better job in identifying ECAV events, the maximum TPR is 68% (coder 3). This suggests that there is room to improve the identification ability of all coders, even those who performed better during this exercise.

Except for coder 3 and coder 9, the point estimates for Precision are always higher than the respective TPR estimates. This suggests that most coders are more prone to produce false negatives than false positives. Because of this, particularly attention should be paid in advising coders 3 and 9 on how to avoid mistakenly identifying events that are not ECAV events (i.e., reduce false positives), while the other coders should get help in improving their ability to identify ECAV events they have missed (i.e., reduce false negatives).

Event Encoding: Reliability

We proceed to examining the reliability and coding validity of ECAV variables encoded from events. For this exercise, coders were provided with 117 identical event descriptions from the CGS and were asked to encode all ECAV variables based on these descriptions.[8] We then assessed the degree of agreement among coders (intercoder reliability) and the ability of each coder to correctly encode the relevant information by comparing it to the Coding Gold Standard (coding validity). We begin with an assessment of reliability. Following the best practices in this field (Neuendorf 2002), intercoder reliability is evaluated globally and individually for each of nine ECAV variables of interest. For each variable, the evaluation is performed by comparing two levels of inter-coder comparison, namely the global and individual level.[9] The global analysis is focused on summarizing the reliability of each variable, whereas the individual level disaggregates reliability information at the coder level, which allow us to examine the relative contribution of each coder to global reliability scores. To conserve space, we present results from the global exercise in the manuscript, while individual analyses are presented in appendix C.

[7] Notice that the random classification threshold (grey line in Figure 1) falls within the TPR confidence intervals for coders 5, 6, 7, 8, 10 and 11. Because of this, we cannot be confident that identification performance of these coders is actually better than a randomized binary classification.

[8] To be precise, we asked coders to encode all variables that require assigning numerical values. For variables where coders assign strings (e.g. location name, event name), calculating reliability and validity scores is not really useful.

[9] For all calculations, categorical variables in ECAV are treated as nominal. While the majority of the variables are expressed on a nominal scale, this is not the case for *Location Precision*, *Participant Number* and *Participant Deaths* which have a clear order among their categories. Analyzing these variables as if they were nominal is not incorrect, but rather is a conservative choice in order to not inflate the results by blurring the boundaries between categories.

Global Reliability

For global comparisons for all 11 coders, we use averaged Cohen's kappa, Fleiss's kappa, and Krippendorff's alpha for more than two coders. The comparisons between pairs of coders are performed using Cohen's kappa, Scott's pi, Krippendorff's alpha for two coders (results in appendix C). These measures are among the most prominent introduced in the reliability literature and are considered better options than more naive Percent Agreement measures. The measures differ on how they are factoring in the expected probability of random agreement among coders (see Appendix C for more details on this aspect). We use several measures to increase transparency of the results presented and to minimize the dependence on a single particular metric. Except for Krippendorff's alpha, all metrics vary from -1 to 1 with -1 representing a level of perfect disagreement, 0 representing agreement no better than chance, and 1 signaling perfect agreement. Krippendorff's alpha varies from 0 to 1, with 1 representing the highest level of agreement.

There is no common standard regarding a good level of agreement, but the literature nevertheless identifies some "rules of thumb." For example, Fleiss (2013) suggests that values of Fleiss's kappa that are lower than 0.40 indicate poor agreement, values from 0.60 to 0.74 signal intermediate to good agreement, and values higher than 0.74 point toward very good agreement. Similarly, discussing Cohen's kappa, Banerjee et al. (1999) evaluate agreement as poor for values below 0.40, fair to good between 0.40 and 0.75, and excellent for values over 0.75. Krippendorff is most conservative and suggests that Krippendorff's alpha values greater than 0.79 indicate good agreement, while only tentative conclusions should be drawn for values between 0.667 and 0.79 (Krippendorff 2004: 241). We thus adopt the following terminology:

- Poor agreement - everything less than 0.40
- Fair to intermediate agreement - values between 0.40 (fair) and 0.60 (intermediate)
- Good agreement - values from 0.61 to 0.74
- Very good agreement - values above 0.74

Table 1 presents the global reliability results by variable. The table shows that nominal values of all measurements are very similar. This is a good sign since an analysis of the aspects driving the differences would have been required otherwise. As can be seen in the table, level of agreement among the 11 coders is *very good* for Event Violence, *good* for Participant Deaths, Event Direction and Actor Type, *fair to intermediate* for Participant Number, Target Type and Actor Side, and *poor* only for Target Side.

Table 1: Intercoder reliability for 11 coders by variable

	Cohen kappa	Fleiss kappa	Krippendorff alpha
Actor.1.Type	0.63	0.63	0.63
Actor.1.Side	0.58	0.58	0.59
Target.1.Type	0.55	0.55	0.55
Target.1.Side	0.39	0.39	0.40
Event.Direction	0.70	0.70	0.69
Event.Violence	0.79	0.79	0.79
Participant.Number	0.44	0.44	0.42
Participant.Deaths	0.72	0.72	0.73
Location.Precision	0.57	0.57	0.59

The relatively lower scores for Actor/Target Type and Side could stem from the greater complexity of these variables since they require a fair amount of interpretation for coding. In addition, Actor/Target type, but also Location Precision consist of more categories than other variables, requiring coders to choose one of seven categories, which will generally produce lower scores. Also interesting is that Target type/side variables have lower scores despite these variables having the same coding structure as Actor variables. We investigate these aspects further in the appendix, where we zoom in and examine agreement for each variable by category. These results show that clearer distinctions between "unknown" and "nonstate actor, citizens" categories could help improve agreement for Actor and Target type variables.

Table 2 compares agreement by variable category, which can help explain why some variables have low overall scores. Table 5 points to a number of aspects of the current coding scheme that could lower agreement among coders. First, the *-99 ("unknown")* category seems be responsible for the lower score of the type variable. Agreement for each category is higher than the overall agreement shown in *Table 1* except for the categories *-99 ("unknown")* and *2 ("nonstate actor, citizens")*. The coding scheme should thus be more explicit in establishing the boundary between *"unknown"* and *"citizens"* categories. Coders may be using these two categories as residual category but not consistently. Similarly, the *-99* category is also responsible for lowering the agreement on the Participant Deaths variable. Moving to the Target variables, *Table 2* shows that Target Type categories *-99 ("unknown"), 1 ("state actor")* and *2 ("nonstate actor, civilians")* suffer from lower agreement. A reason for this is likely that a state actor is often the symbolic target of an event (e.g. a riot) manifesting itself with actions

like the destruction of private properties (e.g. cars and shops). Some coders may choose to code civilians (immediate target) in these cases, while others choose to code the state actor (symbolic target). If this is the case, then the low agreement on all Target variables comes with no surprise.

On the Participant Number variable, coders do not reach good levels of agreement except for events with a very large number of participants. This is striking since the distinction between the categories of this variable is supposedly very clear. Disagreement of the coders is not limited to some categories but involves the entire variable. Clearer guidelines about when and how to code this variable are required if it has to be retained in the coding structure.

The reliability for Location Precision is very good only for the extreme precision levels. At first glance, it seems that the distinction between different levels of administrative units (first-order, second-order) is responsible for poor agreement. However, individual results in the tables below show that most individual scores on this variable are between good and very good, while coder 3 and coder 9 are doing a poor job on this particular variable, which brings down global agreement.

Table 2: Intercoder reliability for 11 coders by variable category (Fleiss's kappa)

Category	Actor.1.Type	Actor.1.Side	Target.1.Type	Target.1.Side	Event.Direction	Event.Violence
-99	0.57	0.55	0.53	0.4		
0		0.71		0.4	0.7	0.79
1	0.78	0.54	0.52	0.36	0.7	0.79
2	0.39		0.42			
3	0.73		0.69			
4	0.83		0.67			
5	0		0.01			
6						

Table 2 (continued)

Category	Participant.Number	Participant.Deaths	Location.Precision
-99	0.43	0.35	
0		0.68	
1	0.37	0.89	0.79
2	-	0.89	0.42
3	0.31	-	0.23
4	-		0.53
5	0.73		0.33
6			0.86

Individual Reliability

Individual analyses allow us to further investigate the relative contribution of each coder to global reliability scores. Table 3 shows a measure of average agreement between each coder and all his colleagues by variable. Table 4 reports the relevant overall reliability by variable to easily compare individual averages and the overall scores. This table is useful to spot coders whose coding differs from the majority of all others.

Table 3: Mean (Scott's pi) individual intercoder reliability by variable

Table 3: Mean (Scott's pi) individual intercoder reliability by variable

	coder1	coder2	coder3	coder4	coder5	coder6	coder7	coder8	coder9	coder10	coder11
Actor.1.Type	0.65	0.54	0.6	0.68	0.63	0.69	0.69	0.52	0.6	0.66	0.67
Actor.1.Side	0.47	0.51	0.66	0.66	0.59	0.46	0.54	0.54	0.62	0.68	0.65
Target.1.Type	0.53	0.59	0.54	0.56	0.5	0.48	0.62	0.45	0.54	0.58	0.61
Target.1.Side	0.3	0.26	0.46	0.44	0.42	0.27	0.39	0.28	0.4	0.4	0.51
Event.Direction	0.69	0.76	0.77	0.56	0.76	0.58	0.75	0.55	0.76	0.72	0.66
Event.Violence	0.79	0.79	0.64	0.78	0.82	0.83	0.8	0.75	0.83	0.77	0.84
Participant.Number	0.5	0.47	0.39	0.44	0.49	0.26	0.46	0.25	0.32	0.49	0.44
Participant.Deaths	0.73	0.76	0.67	0.73	0.73	0.67	0.73	0.72	0.71	0.76	0.72
Location.Precision	0.68	0.66	0.32	0.64	0.53	0.67	0.65	0.65	0.31	0.64	0.52

Table 4: Overall Fleiss's kappa from Table 1

	Fleiss kappa
Actor.1.Type	0.63
Actor.1.Side	0.58
Target.1.Type	0.55
Target.1.Side	0.39
Event.Direction	0.70
Event.Violence	0.79
Participant.Number	0.44
Participant.Deaths	0.72
Location.Precision	0.57

As shown in Table 3, coders 4, 6 and 8 have a low average agreement with their colleagues on the Event Direction variable. In particular, coder 8 has an average agreement that is always below the overall agreement on the first 5 variables in table C3. The exercise also reveals that Participant Number is characterized by very low reliability scores for coders 6, 8 and 9. Unfortunately, the already discussed low overall agreement for Target Type and Target Side seems to be the result of a diffused situation of disagreement.

Event Encoding: Validity

Global Validity

We proceed to an assessment of coding validity, meaning that we compare the variable coding of all coders to the Coding Gold Standard (CGS). We also assess whether validity issues align with reliability issues or not. Table 5 allows to make this comparison on the Global level by presenting the overall reliability and coding validity results side by side.

Table 5: Global intercoder reliability and coding validity (Fleiss's kappa)

	Intercoder Reliability	Coding Validity
Actor.1.Type	0.63	0.71
Actor.1.Side	0.58	0.49
Target.1.Type	0.55	0.56
Target.1.Side	0.39	0.34
Event.Direction	0.70	0.56
Event.Violence	0.79	0.82
Participant.Number	0.44	0.42
Participant.Deaths	0.72	0.75
Location.Precision	0.57	0.71

According to Table 5, overall coding validity is *very good* on Event Violence and Participant Deaths, *good* for Actor Type and Location Precision, *fair to intermediate* for Event Direction, Participant Number, Target Type and Actor Side, and *poor* only for Target Side. This picture of coding validity is quite similar to the one for intercoder reliability, with some exceptions. Event Direction reaches a good level of reliability (.70), but only a fair-intermediate level of validity (.56), and Actor Side moves from an intermediate level of reliability (.58) to a fair level of validity (.49). On the other hand, the Location Precision score is way better on validity (.71) than it is on reliability (.57). In the appendix, we also compare each individual coder to the GCS.

Individual Validity

We proceed to a discussion of coder validity at the individual level. In the manuscript, we present global validity results, while we show the contribution of individual coders by comparing their coding to the GCS here.

Table 6: Agreement between each coder and the Coded Gold Standard (Scott's pi)

Table 6: Agreement between each coder and the Coded Gold Standard (Scott's pi)

	coder1	coder2	coder3	coder4	coder5	coder6	coder7	coder8	coder9	coder10	coder11
Actor.1.Type	0.73	0.69	0.62	0.76	0.65	0.85	0.79	0.52	0.65	0.81	0.78
Actor.1.Side	0.30	0.46	0.49	0.48	0.47	0.71	0.40	0.44	0.49	0.61	0.49
Target.1.Type	0.50	0.58	0.52	0.49	0.39	0.69	0.61	0.54	0.61	0.54	0.67
Target.1.Side	0.14	0.12	0.24	0.24	0.33	0.62	0.38	0.42	0.30	0.46	0.44
Event.Direction	0.47	0.61	0.53	0.73	0.51	0.82	0.54	0.26	0.64	0.54	0.53
Event.Violence	0.76	0.88	0.62	0.82	0.87	0.89	0.82	0.80	0.89	0.79	0.89
Participant.Number	0.40	0.29	0.55	0.31	0.47	0.17	0.73	0.15	0.57	0.47	0.54
Participant.Deaths	0.66	0.74	0.89	0.65	0.94	0.57	0.65	0.92	0.91	0.73	0.64
Location.Precision	0.91	0.91	0.32	0.81	0.65	0.87	0.81	0.81	0.31	0.79	0.63

Table 6 shows that coder 6 has the highest levels of agreement with the gold standard except for the variables Participant Number (.17) and Participant Deaths (.57). Coder 8 has poor validity performances. Different to what emerged from the reliability assessment, coders seem to overall do a good job in coding the Location Precision variable. The low value for location precision is the result of poor performance of coder 3 and coder 9 on this variable.

Conclusion

The assessments of event identification and encoding show an encouraging picture with good and consistent performances in both assessments. The analyses show how in-depth analyses of coding quality can be useful for several reasons. First, and most importantly, they help ensure data quality. While this paper describes the assessment of coding quality among the initial coders recruited for ECAV, subsequent exercises were used throughout the project. Second, such an exercise can be useful to clarify coding procedures. Results from this exercise helped improve the clarity of the coding scheme especially with regard to actor type variables, participant numbers. Third, assessments can and should be used to guide the selection of high-quality coders, since the identification and encoding decisions by individual coders did have a significant impact on overall scores in both assessments. We provided additional support for coders with lower scores and used results to inform our decisions on coder retention.

References

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. "Beyond Kappa: A Review of Interrater Agreement Measures." *Canadian Journal of Statistics* 27 (1):3–23.

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 (4):382–401. https://doi.org/10.1080/03050629.2012.696981.

Daxecker, Ursula E., Elio Amicarelli, and Alexander Jung. 2019. "Electoral Contention and Violence (ECAV): A New Dataset." *Journal of Peace Research.*

Fleiss, Joseph L., Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Höglund, Kristine, and Magnus Oberg. 2011. *Understanding Peace Research: Methods and Challenges*. Taylor & Francis. http://books.google.nl/books?hl=nl&lr=&id=rxsRitHQbGEC&oi=fnd&pg=PP1&dq=Understanding+Peace+Research:+Methods+and+Challenges&ots=RlCuWo96Ag&sig=ppyslowVkxq8KBcihVAIDBq9bUs.

Neuendorf, Kimberly A. 2002. *The Content Analysis Guidebook*. Sage.

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 (3):340–61. https://doi.org/10.1080/03050629.2011.596028.

Urdal, Henrik. 2008. "Urban Social Disturbance in Africa and Asia." Oslo: PRIO.

Weidmann, Nils B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60 (1):206–18. https://doi.org/10.1111/ajps.12196.