Electoral Contention and Violence (ECAV) Data Project
Assessment of intercoder reliability and coding validity

Elio Amicarelli

doc ref: ECAV10016

1. *Introduction*

This report concludes the evaluation of the ECAV data quality by examining the intercoder reliability and coding validity of the ECAV variables. The assessment is performed in accordance with the strategy outlined in the document reference ECAV08016 (section 2.3).

Following the best practices in this field (Neuendorff 2002), intercoder reliability is evaluated individually for each of 9 ECAV variables of interest. For each variable, this evaluation is performed by looking at two different levels of inter-coder comparison, namely the global and individual level. While the global level analysis is focused on providing a unique index summarizing the degrees of reliability for each variable, at the individual level this information is disaggregated at the coder level. Thus this enables the Project Leader to examine the relative contribution of each coder to the global reliability scores.

The ability of each coder to produce valid data is examined by comparing his coding output against a Coded Gold Standard.

2. *Setting and Methods*

2.1 Setting

The general setting for this task has been outlined in ECAV08016 and involves the following steps:

1. Let E be a set of n ECAV events' descriptions the coders have never coded before. Create a Coded Gold Standard $CGS = (y_1, y_2, y_3, ..., y_n)$ where y_i is the correct ECAV event encoded in 9 ECAV variables from the information contained in a given element of E. *Status: Completed*

2. Let $C_i = (\hat{y}_{i1}, \hat{y}_{i2}, \hat{y}_{i3}, ..., \hat{y}_{in})$ be the set of n events coded in 9 ECAV variables by a general *coder* i over the set of n descriptions E. Obtain C_i for each coder currently hired so that the collection of all coders' outputs is defined as C. *Status: Completed*

3. Assess the intercoder reliability. Define C^v as the the collection of all coders' output for a general ECAV variable v. For each ECAV variable, the *global intercoder reliability* is assessed by applying a comparison function with the following general form:

   $$f(C^v)$$

   For each ECAV variable, the *individual intercoder reliability* is assessed by applying for each pair of different coders i and j a function with the following general form:

   $$f(C_i^v, C_j^v)$$

   Staus: Finalized in this report

4. Assess the coding validity. For each ECAV variable, the *coding validity* is assessed by applying for each coder a function with the following general form:

   $$f(CGS^v, C_i^v)$$

   Staus: Finalized in this report

2.2 Methods

In this section I describe the statistical measures used in this report to assess intercoder reliability and coding validity. The section is mainly focused on the motivation behind the selection of this measurement devices and their interpretation. For mathematical details please see the separate Appendix.

2.2.1 Metrics adopted

This report contains two type of comparisons, namely comparisons performed between pairs of coders (section 3.1.2 and section 3.2) and comparisons performed between more than two coders (section 3.1.1 and 3.2). The comparisons between pairs of coders are performed using

a) Cohen's kappa

b) Scott's pi

c) Krippendorff's alpha for two coders

In order to make comparisons among the 11 coders at once I use

d) averaged Cohen's kappa

e) Fleiss's kappa

f) Krippendorff's alpha for more than two coders

These measures are among the most prominent introduced in the reliability literature. All of them are considered better options than excessively naive Percent Agreement measure (not adopted in this report), but at the same time they differ on how they are factoring in the expected probability of random agreement among coders (see Appendix for more details on this aspect).

It is important to notice that the use of different measures is dictated by the need of increasing the level of transparency of the results presented and minimizing the dependence on a single particular metric. While on one hand the need of considering different perspectives is addressed by using more than one metric, on the other hand the simultaneous examination of the results regarding pairs of coders with those regarding all of them is ensured by the fact that *d)* is a generalization of *a)*, *e)* is a generalization of *b)* and *f)* is a generalization of *c)*.

Another important aspect of the analysis presented in the next section is that all the categorical variables 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 wrong, instead it is a conservative choice I made in order to not inflate the results by blurring the boundaries between categories.

2.2.2 Guidelines for the interpretation of results

All the metrics adopted except for Krippendorff's alpha 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 are no common standards regarding what has to be considered a good level of agreement. Some of the rules of thumb proposed for "kappa like" statistics have some overlaps, though. For example, Fleiss (1981) suggests that Fleiss's kappa less than 0.40 indicate poor agreement, values from 0.60 to 0.74 signal intermediate to good agreement and values bigger 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 has one of the most conservative approaches and he suggests that Krippendorff's alpha values greater than 0.79 indicate good agreement while "tentative conclusions are still acceptable" between 0.667 and 0.79 (Krippendorff 2004).

For what concerns Cohen's Kappa, Fleiss's Kappa and Scott's pi, this study adopts 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

3. *Results*

3.1.1 Intercoder reliability - Global level

Table 1 shows the global reliability results by variable. As can be seen, setting aside issues regarding the interpretation of the different metrics, their nominal values are all extremely similar. This is a good sign. In case of striking differences among them an analysis of the aspects driving the differences would have been required.

According with *Table 1*, the level of agreement among the 11 coders is *very good* on 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

	avg 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

Based on the ECAV codebook (Daxecker and Amicarelli, 2014), I would argue that Actor/Target Type and Side can be considered as more complex variables if compared with all the others as they require a relevant amount of interpretation in order to be coded. In light of this statement, it is interesting to explore why coders do not reach good or very good level of agreement on Participant Number and Location Precision which are supposed to be easy to code. A second aspect worth of further consideration is the difference between Actor-related and Target-related variables: despite these variables have the same coding structure, agreement on Target variables is lower than agreement on Actor variables. These and other aspects can be investigated by zooming in and looking at how the agreement for each variable is distributed among its categories. This type of information is presented in *Table 2*.

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

The results showed in *Table 2* are very interesting because they point out to possible aspects of the current coding rules that are negatively affecting the agreement among coders.

First of all, the *-99 ("unknown")* category seems to be responsible for lowering the agreement on some variables. This seems to be the case for Actor Type, where the agreement for each category is higher than the overall agreement showed in *Table 1* except for the categories *-99 ("unknown")* and *2 ("nonstate actor, civilians")*. These results suggest that given the current formulation of the coding rules, the main issue with the Actor Type variable may be that the boundary between the *"unknown"* and *"civilians"* categories is not clear enough. Because of this, it could be the case that the coders differ in how they are using one of these two categories as residual category. Similarly, the low agreement on the *-99* category is also responsible for lowering the agreement on the Participant Deaths variable.

Moving to the Target variables, *Table 2* shows that for Type categories *-99("unknown"), 1 ("state actor")* and *2 ("nonstate actor, civilians")* suffer from lower agreement. It is often the case that a state actor is 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). In similar cases it is possible that the coders are confused about how to code the Target Type variable so that some of them choose to code civilians (immediate target) while others choose to code the state actor (symbolic target). If this is the case, then the low agreement on all the Target Side categories comes with no surprise.

On the Participant Number variable the coders do not reach good levels of agreement except for those events with a very large number of participants. This is striking since the distinction between the categories of this variable is supposedly very clear. The results point toward the fact that the disagreement of the coders is not limited to some categories but involves the entire variable! Probably more clear 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 it seems that the distinction among intermediate administrative units is responsible for generating confusion among coders.

However, by looking at *Table 3* and *Table 6*, it could be seen that the individual scores are all between good and very good except for coder 3 and coder 9 who are doing a terrible job on this particular variable.

3.1.2 Intercoder reliability - Individual level

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 so that it is possible to easily eyeball between the individual averages and the overall scores. This table is useful to spot those coders who are coding a given variable differently than the majority of all others.

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 showed in *Table 3* coders 4, 6 and 8 have a low average agreement with their colleagues on the Event Direction variable. In particular, on the first 5 variables coder 8 has an average agreement which is always below the overall agreement presented in *Table 4*. Unfortunately, the low overall agreement already discussed for Target Type and Target Side seems to be the result of a diffused situation of disagreement. On the contrary, Participant Number is characterized by very low reliability scores for coders 6, 8 and 9.

3.2 Coding validity

I now move to the assessment of coding validity. Since reliability and validity are two very different concepts (see ECAV08016), at this stage of the analysis it is useful to compare the main patterns already identified during the reliability analysis with the validity figures. In other words, it is useful to 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 with *Table 5*, the overall level of 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 regarding intercoder reliability except for Event Direction which reaches a good level of reliability (.70) but only a fair-intermediate level of validity (.56) and for Actor Side which moved 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).

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

According with *Table 6*, coder 6 is the one with the highest levels of agreement with the gold standard except for the variables Participant Number (.17) and Participant Deaths (.57). As well as for the individual reliability, coder 8 has poor validity performances. Contrary to what emerged from the reliability assessment, it seems that overall the coders are doing a good job in coding the Location Precision variable correctly. It is clear at this point that, the negative impact of coder 3 and coder 9 is the main driver of the not-so good results previously examined on this variable.

4. *Recommendations*

In this report I examined the quality of the ECAV data under different perspectives. The results show an encouraging picture with good and consistent performances in both the reliability and validity assessment. Indeed, there is room to improve the quality of the ECAV data especially by fixing some aspects of the coding rules which revealed to be potentially confusing for the coders.

In this regards, the Project Leader may consider to improve the specification and training on the following coding aspects:

- Better define the rules governing the coders choice to resort to the *-99* category for those variables admitting this value.

- Better define the instances where the Target of an event should be coded based on either the symbolic or immediate target.
- Better define the type of events where the coders should code information for the Participant Number variable and the level of subjective interpretation they are allowed to exert on it.

Additionally, it is recommended to repeat the identification and reliability/validity exercises every time a new set of coders is hired. It would be beneficial to carry out these exercises after the coders have received their coding training and before they start the actual coding. By using the material and the analytical structure produced during this pilot study, the assessment phase should take approximately two weeks of full-time work. By considering together the identification and reliability/validity performances, the Project Leader may decide to further enhance the expected quality of the data by retaining only the coders achieving satisfying performances.

References

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27(1), 3-23.*

Daxecker, U. E., and Amicarelli, E. (2014). Electoral Contention and Violence (ECAV): A New Dataset. *Paper presented at the European Conflict Network meeting, October 16-18, 2014, Uppsala, Sweden.*

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Thousand Oaks, California: Sage.

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, California: Sage Publications.