The Militarized Interstate Confrontation (MIC) Codebook

This document provides user information on the coding procedures for the Militarized Interstate Confrontation (MICs), 1816-2014. We include a discussion of data formats and variable information for the confrontation-level data. We also discuss the addition of new variables for the data, including `fatalmin` and `fatalmax`. Any questions or errors should be reported to the authors.

July 11, 2023

Citation

We ask that users of this dataset cite the following article and note the version number of the data they are using in their study:

- D. M. Gibler and S. V. Miller, 'The Militarized Interstate Confrontation (MIC) Dataset, 1816-2014." *Journal of Conflict Resolution*, 2023. doi:10.1177/00220027221104704

Rationale for Militarized Interstate Confrontations (MIC) Dataset

Our data collection efforts began with the original Militarized Interstate Dispute (MID) dataset from the Correlates of War Project (CoWMID) (Jones, Bremer and Singer, 1996). We found an overwhelming number of errors in the original data and corrected those data, providing all of our suggested changes to the CoWMID Project dataset hosts. However, after more than nine years of communications with CoWMID and several exchanges in *International Studies Quarterly*,[1] it is apparent that CoWMID admits their pre-2002 data is incredibly error-prone but that they have neither the resources nor inclination to fix those thousands of admitted errors. Therefore, we are introducing the Militarized Interstate Confrontation (MICs) data, which are based on proper use of MID coding rules and include numerous advancements in both information and presentation.[2]

We use this document to discuss how we coded the data, using original CoWMID coding documents, and we provide annotations on specific variables most often used (and, when appropriate, misused) by conflict scholars. We also discuss important new variables we have added to the dataset and several variables we remove from the data.

Data Files Associated with the Militarized Interstate Confrontation Data

We provide the following files for users of the MIC data set. We advertise the data of interest in the list below first as the more accessible comma-separated values file (`.csv`), but we also have the same data available in a Stata data file format (`.dta`) or an R serialized data frame (`.rds`). Please note that we are constantly revising these data; users should report the version number of the dataset used in any research. Future releases of the data will come with a text file summarizing changes to the data.

- `mic-conf-[version].csv`: This is the confrontation-level MIC data with one case per confrontation. We also include Stata and R versions of the data.

Two Levels of Data and the Variables in Each Dataset

We keep the confrontation-level and participant-level presentation of the data and faithfully code values for each variable according to the rules established in Jones, Bremer and Singer (1996) and amended by Ghosn, Palmer and Bremer (2004) and Palmer et al. (2015). Our separate

---

[1] See Gibler, Miller and Little (2016) and Gibler, Miller and Little (2020) for examples, but also our final response which is posted on this website in Gibler and Miller (N.d.). Note that our final response demonstrates that the CoWMID hosts (1) were ignorant of basic coding rules that led to the creation of the dataset—coding rules that should still be in use today– and (2) admitted there are thousands of errors in their released data.

[2] We previously released these suggested changes to CoWMIS and the public in 2013 and have amended these data over time. CoWMID has referred to our suggestions as the GML data based on the author listing of our first *ISQ* piece, but this is the first time we are releasing the cases as a coherent dataset, and this is our first release of a codebook.

change, drop, merge, and could-not-find reports provide details on conflict cases in which we disagree with the CoWMID team's application of their own coding rules as established in prior publications. Here, we describe the variables we include and exclude from the data and also describe several changes we made in order to better facilitate studies of conflict.

MIC Confrontation-Level Data

The following variables are included in the MIC confrontation-level data, with cautionary notes about each variable when appropriate:

- `micnum`: The confrontation number for each case, as originally established by CoWMID for each case. Note that we have also added numerous confrontation cases that are not in the CoWMID data as disputes. These cases begin with `micnum` values in the 9000s to demonstrate completely new cases.

- `stmon`: The start month of the confrontation, with values ranging from 1 to 12. There are no missing month values. If the month of a militarized action could not be determined, the action was not coded.

- `stday`: The start day of the confrontation, with values ranging from 1 to 31 and missing days reported as values of -9. (Note that we have done our best to appeal to the historical record to eliminate missing days in the data, but there are numerous cases where only the month of the militarized action is known.)

- `styear`: The start year of the confrontation, with values ranging from 1816 to 2014. There are no missing year values. If the year of a militarized action could not be determined, the action was not coded.

- `endmon`: The end month of the confrontation, with values ranging from 1 to 12. Note that the end date of a confrontation can be coded due to the last incident, a set period of time following certain incidents (e.g. a blockade that last six months), or the implementation of a treaty (see especially, Jones, Bremer and Singer, 1996). For the last militarized event in a confrontation, please see our Militarized Interstate Events (MIE) data, 1816-2014. We also provide a MICEndings dataset that codes what ended the confrontation if the end date does not correspond to a militarized event. MICEndings includes data on whether a treaty was signed, provisions of the treaty or agreement, etc.

- `endday`: The end day of the confrontation, with values ranging from 1 to 31 and missing days reported as -9. (Again, we have done our best to appeal to the historical record to eliminate missing days in the data, but there are numerous cases where only the end month of the militarized action is known.)

- `endyear`: The end year of the confrontation, with values ranging from 1816 to 2014.

- `fatalmin`: This is an estimate of the minimum number of military battle-deaths in the confrontation. The CoWMID project often had to code missing values for `fatality` because the `fatalpre` was unknown but near one of the battle-death thresholds. For example, if military battle-deaths were likely, between 20 and 30, but not definitively 25 or less or 26 or more, a missing `fatality` was coded because it was unclear which value of the scale should be entered. We change this by introducing lower- and upper-bound estimates of `fatalmin` and `fatalmax`. The lower-bound estimate also allows researchers to determine

confrontations with military battle-deaths (or fatal MICs) since we now have no missing values for this variable. Finally, by including the fatality range variables rather than the fatality scale, we are now able to integrate fatality estimates from available war data (see the section on the integrated war data provided below).

- `fatalmax`: This is an estimate of the maximum number of military battle-deaths in the confrontation. Used in conjunction with `fatalmin`, the most likely range of military battle-deaths in the confrontation can be determined.

- `hiact`: The "highest" level of militarized action by any state in the confrontation. Note that this scale is actually not ordinal, however. Possible values [with `hostlev` in brackets] include the following: 0 No militarized action [1] 1 Threat to use force [2], 2 Threat to blockade [2], 3 Threat to occupy territory [2], 4 Threat to declare war [2], 5 Threat to use CBR weapons [2], 6 Threat to join war, 7 Show of force [3], 8 Alert [3], 9 Nuclear alert [3], 10 Mobilization [3], 11 Fortify border [3], 12 Border violation [3], 13 Blockade [4], 14 Occupation of territory [4], 15 Seizure [4], 16 Attack [4], 17 Clash [4], 18 Declaration of war [4] 19 Use of CBR weapons [4], and 22 War battle [5].[3]

- `hostlev`: The "highest" recorded hostility level by any state in the confrontation. Again, this scale should not be treated as ordinal. Possible values include the following: 1 No militarized action, 2 Threat to use force, 3 Display of force, 4 Use of force, and 5 War.

- `outcome`: This variable describes how the confrontation ended and can have a value of 1 Victory for side A, 2 Victory for side B, 3 Yield by side A, 4 Yield by side B, 5 Stalemate, 6 Compromise, 7 Released, 8 Unclear, 9 Joins ongoing war. Please see our notes below before using this variable.

- `numa`: The number of participant states on Side A of the confrontation, with Side A defined by whichever state initiated the first militarized action in the confrontation. All states coordinating with the original Side A state are added to the number value in `numa`. Note that Side A is the lower-number ccode for confrontations that began with clashes (confrontations in which the initiator could not be determined). Also note that states that enter and leave the confrontation over time do not count as additional participants. For example, Syria enters and leaves the 1993-2014 confrontation between Lebanon and Israel several times because there are six-month gaps between militarized events (see MIC#4182).

- `numb`: The number of participants on Side B of the confrontation, with coordinated fighting against Side A. States that enter and leave the confrontation over time do not count as additional participants.

- `version`: The MIC confrontation-level data version number. We are constantly reviewing and updating our datasets, and we ask users of our data to always report which version number of the data they are using in their research.

---

[3]Please see our posted response for clarification of several types of systematic errors in CoWMID decisions to include nearly 100 disputes in their data (Gibler and Miller, N.d.). For example, we do not remain agnostic to contemporaneous coverage of events and code propaganda. We do not code shows of force against civilians as this was never allowed in CoWMID coding rules. Finally, we continue to distinguish between attacks and clashes based on whether the initiator could be determined. If there is a clear initiator, the event is an attack; with no clear initiator, we code the event as a clash, as per the original CoWMID coding rules.

A Note on Outcomes and Settlements

We are currently collecting issue-level data for all issues that were militarized in the MIC dataset. This issue-level data will eventually lead to the removal of the outcome variable from the base confrontation data set. We are collecting new data because the coding rules for these variables were not consistently applied in the MID data. As we point out in the data release manuscript, the coding of war outcomes was particularly inconsistent. The outcome variable value of "joins ongoing war" should be used for many of the confrontation cases that precede larger wars, and the proper highest action and hostility level in the dispute should be recorded. Numerous MIDs in the original data had only one state within the MID joining a war, which rarely makes sense logically; again, changing the outcome variable value to "joins ongoing war" should be used for these cases. We correct all of these errors in the MIC data.

Value codes of yield, released, and unclear for outcomes are inconsistently applied in the CoWMID data, and the differences in practice between negotiated, none, and unclear settlements are not clear. Further, as CoWMID moved to newspaper-based sourcing, the emphasis on the dispute-level endings has waned. MID4 was collected using news stories that described the onset of possible disputes. According to the bibliography for the data, no effort was made to use secondary sources to determine outcomes and settlements. No search terms were used for settlement signings and the like, which suggests the main concern was onset and type of conflict. This is problematic for these two variables. Victories are four times less likely to be coded, and yields are ten times less likely in MID4. Stalemates are one and a half times more likely in MID4 while compromises and released disputes are five and three times less likely to be found in MID4, respectively. Among settlements, missing values were coded at a rate ten times higher in MID4 than in previous versions of the MID data, while negotiated and imposed settlements were eight and four times less likely to be found, respectively. Unclear endings are about 50% more likely to be coded in MID4. The new CoWMID procedures for collecting dispute sources makes it difficult to code these variables, and any values for this variable in the CoWMID data seem to be based on poor (if any) empirical evidence.

We have made changes to many of the values of these variables in the MIC data, but users are encouraged to proceed with caution when examining outcomes. We have not performed the kind of systematic review we conducted on variables such as participants, dates, and actions. Instead, rather than repeat the original coding rules that were difficult to interpret, we are developing the new dataset of issues that documents change in status quo issue positions over time and state-level participation in bargaining over those issues. That data will be easily integrated with the MIC and MIE datasets.

Variables Not Included in the Militarized Interstate Confrontation-Level Data

We omit the following variables, with explanations following each variable definition.

- `fatality`: We drop the seven-category CoWMID measure of fatalities. Researchers can still calculate their own ordinal fatality categories based on the information available in `fatalmin` and `fatalmax`. However, we believe studies will be better served by continuous measures of fatalities (that are perhaps logged).

- `fatalpre`: CoW's fatality estimates, with precision, are routinely unknown if they are not zero. We instead use the `fatalmin` and `fatalmax` categories to communicate this information and, with it, uncertainty around estimated fatalities in confrontations. Researchers can

calculate their own version of this fatalities (with precision) variable. Cases where `fatalmin` equals `fatalmax` are cases of known fatalities, with precision.

- `recip`: Gibler, Miller and Little (2016) first noted that dispute-level reciprocation is one of the more haphazardly coded and least understood variables in the CoWMID data set, and that reliance on this variable could lead to false inferences. Used primarily as a proxy for audience costs analyses using dispute-level data, the variable only minimally informs the researcher whether a participant on Side B initiated a militarized incident. Multiple issues arise from a researcher's temptation to interpret this variable as anything more than what it minimally is. Consider a single-incident dispute in which Side A attacked Side B in a surprise ambush, but Side B fought off the attack with deadly force. In this simple case, Side B responded to Side A's attack with deadly force, but the clear initiation by Side A means Side B would be coded as having no militarized incident (and, thus, the reciprocation variable would be 0) even though Side B responded in real time with deadly force. Consider a near identical scenario, but the remoteness of the ambush (either spatially or temporally) means our exhaustive archival research could not indicate who initiated the ambush. Under those conditions, both sides are coded as engaging in a clash and the reciprocation variable would be 1 by default. The reciprocation variable is not nearly as informative as researchers may want it to be. Thus, we remove this variable from our data set and invite the researcher to recreate a version of it, if they like, based on our confrontation-participant data (or our forthcoming events data). If a state on side B of a confrontation has an event, reciprocation can be inferred to have occurred according to Jones, Bremer and Singer (1996) coding rules.

- `settle`: Our review of the data, along with CoWMID's developing approaches to data collection, suggest this variable is no longer worth maintaining as a core component of a base data set on confrontations. Indeed, we have found almost 400 cases which were improperly coded. Either CoWMID grafted a settlement code to a dispute when the settlement occurred years after the dispute, or CoWMID may have just missed an actual settlement that ended the dispute. Further, the typology of "imposed" and "negotiated" belies substantial heterogeneity in what exactly is settling a confrontation episode. A "settlement" can be anything from a resort to the International Court of Justice to an exchange of letters. Further, the data-generating process behind CoWMID is increasingly ignoring any types of conflict resolution. MID4 was collected using news stories that described the onset of possible disputes. According to the bibliography for the data, no effort was made to use secondary sources to determine settlements (and outcomes). No search terms were used for settlement signings and the like, which suggests the main concern was onset and type of conflict. Under those conditions, we see no real value in maintaining the `settle` variable as part of the base confrontation data set. Instead, we refer the researcher to the MICEndings data set we provide for more granular information on settlement information for the temporal domain of 1816 to 2014.

- `dispnum4`: The MID4 release of the CoWMID data provided a new dispute number. All values are missing for pre-1993 cases. Our focus is on the entire domain of confrontation data, 1816-2014, so we see no need to include this variable in our data.

- `maxdur`: The maximum duration of the confrontation can easily be calculated by interested researchers based on the beginning and end dates we provide. More importantly, with many missing days in both the `stday` and `endday`, we encourage those interested in duration to consider carefully their modeling choices. Missing days can be at the start, middle, or

end of the month. In most cases we have narrowed when within the month an event occurred (see the MIE data), but this is still a research design issue.

- `mindur`: The minimum duration of the confrontation can easily be calculated by interested researchers based on the beginning and end dates we provide. As with `maxdur` above, we encourage researchers to consider carefully their modeling choices for confrontation durations.

- `link1`, `link2`, and `link3`: The links variables suggest relationships across cases, including wars, that were developed as part of MID3 and especially MID4. These have never been been developed for cases prior to 1993. Since we have fully integrated the war data into the MIC data, and since we are currently completing an issues-based dataset of all militarized issues, we do not include these variables in our data.

A Discussion of Coding Differences with the CoWMID data

CoWMID's last response to our suggested changes highlighted well several differences in how CoWMID coding rules are interpreted. We briefly reiterate those differences here:

1. Perhaps the biggest misunderstanding concerns the difference between clashes and attacks. CoWMID coding manuals are ambiguous when defining clashes, so we referenced Gochman and Maoz (1984, 589), the original article that introduced the dataset, and used their unambiguous definition of a clash as military hostilities in which the initiator could not be determined. These differ from attacks, which are coded as hostilities with a clear initiator. The CoWMID datahosts wrote that they were ignorant of this original definition; this suggests that CoWMID has substantially variation in how the majority of actions in its dataset are coded.

2. We did not code propaganda. If reports about militarized events were ambiguous, we consulted multiple sources and assessments of the information, including secondary sources and reviews.

3. We did not code shows of force against private citizens. These make no sense in light of the clear coding rules established in Jones, Bremer and Singer (1996) and their subsequent emendations.

4. We did not code cases for which we have no information. CoWMID continues to include 19 cases for which no source information exists. We did not code phantom cases of conflict.

5. We did not code border closures as blockades. A border closure denies entry into the initiator's own state; it does not necessarily deny the entry of goods and citizens into other states.

Treatment of the Conflicts Leading to the World Wars

There has always been an underlying tension between how the CoWMID and the CoWWar datasets are coded. The datasets often do not agree on case values, including fatalities. As Sarkees and Wayman (2010) often imply in the "coding notes" sections of their case studies, it is difficult to merge the two datasets. Multiple disputes are sometimes associated with wars, or disputes precede but do not temporally overlap the war cases. This is strange if we assume, as most do, that wars are a subset of disputes with military fatalities greater than 1,000.

Jones, Bremer and Singer (1996, 176) include the following language when discussing wars:

> When a dispute ends up in war, we treat the participants somewhat differently with regard to the aggregation of incidents. When two states go to war, all other ongoing disputes between these two states cease. Any dispute that erupts between a war belligerent and a non-belligerent state is treated as a separate dispute and only merged with the "core" war if the non-belligerent actively joins the war. If the entry into an ongoing war occurs within six months of its onset, then all sub-war militarized actions between a war belligerent and the third party entrant will be considered as part of incidents leading up to the intervention into the war. In cases when war intervention occurs six months or more after the start of the war, a separate militarized interstate dispute exists between the war belligerent and the other state up to its official entry in the ongoing war; thereafter, all actions are coded as part of the ongoing war. A state can be a participant in a war at a lower level of hostility only if its actions are fully coordinated with the war participants and its military combat falls short of the war threshold.

These coding rules are rarely followed by CoWMID. For example, MID#3702 is a bilateral dispute between Germany and Belgium that begins on November 5, 1939 and ends on May 10, 1940, with the German invasion. Belgium's highest action is an attack while Germany is coded as joining an ongoing war. Meanwhile, World War II (MID#0258) includes Belgian participation from August 25, 1939 until it leaves the system after German occupation on May 28, 1940. CoWMID codes Belgium's and Germany's highest actions as beginning an interstate war. These disputes effectively double count all the militarized incidents between Belgium and Germany during that period.

In another case, Luxembourg is coded as having a two-day, bilateral dispute with Germany, beginning on May 9, 1940 and ending with German occupation. Britain and France are excluded from the dispute even though France had more than 15,000 troops in the conflict, and the British RAF involved one squadron. Five French soldiers and one British soldier were killed in the invasion, but these are also ignored by CoWMID's coding of the case.

Perhaps more important than the inconsistency in coding rule application is the fact that the rules for wars differ conceptually between the CoWMID and CoWWar data. Italian participation in World War II is a case that highlights some of this conceptual confusion as a war is treated differently in the CoWMID data. CoWMID codes a separate dispute for Italian actions against France, beginning on April 15, 1940. The dispute ends with an Italian declaration of war on June 10, 1940, and its highest action is coded as joins ongoing war. We could find no actions on April 15th, but we did find an Italian show of force against French and British forces on April 17th. Since Italy had started coordinating with Germany this show of force was Italy's entry into the larger conflict. CoWMID instead treats this action as distinct from the larger conflict, using the declaration of war to artificially create more conflicts in the data. Wars should simply be a subset of disputes with the highest fatalities—indeed, this was the *raison d'être* for collecting the original CoWMID data—, but instead the presence of war, especially the world wars, alters the data generation process and substantially inflates the number of conflict cases.

We change this type of treatment of the war cases and provide a unified coding of conflict cases in our data. We return to the base Jones, Bremer and Singer (1996) coding rules and define a confrontation as any threat, display, or use of force between two or more states. We define originator(s) on side A of the confrontation as the state that initiated the first militarized event, and, for multilateral conflicts, we identify parties to the confrontation as those states actively fighting together and/or coordinating their participation. We do not try to define when

a confrontation becomes a war. However, with our events data, researchers can easily examine how conflicts unfold and, if necessary, define for themselves when escalation to war occurs—by fatality thresholds, actions, or another indicator.

This unification of conflict cases leads us to merge several conflicts with the world wars. We merge Belgian (#3701), Dutch (#3702), Luxembourgish (#3703), American (#0339 and #0414), Russian (#3822), and Thai (#1694) participation in World War II with the core militarized confrontation (#0258). We do not merge cases in which the parties were not coordinating actions or fighting together, which describes such cases as the Soviet takeovers of Lithuania (#0504), Latvia (#0022), and Estonia (#0505), or the British actions against Bulgaria (#0518) prior to Bulgaria's decision to join the Axis powers in 1941. Bulgaria was not yet coordinating with the Germans.

Reliability, Validity, and Robustness Checks

We agree wholeheartedly with the CoWMID project that, with the exception of protest-dependent cases, MIDs are an excellent concept for identifying interstate conflicts. Unfortunately, we found consistent problems with the reliability of the application of those coding rules to widely available source information. That is why our dataset now exists.

Finally, a seemingly natural implication of the back-and-forth with CoWMID and our introduction of the MIC data is that both datasets can be used as robustness checks for empirical research. However, given what we know of the data and what the CoWMID datahosts have demonstrated, we think that use of the MID data for robustness checks is possible only for confrontations between 2002 and 2014. The error rate in prior versions of the CoWMID data is just too high (a lower bound of 2,588 errors at the dispute level and 5,913 at the participant level derived from CoWMID estimates). So, for example, consider a hypothetical in which estimates of the independent variable of interest are statistically significant using only our data. What does this mean? Does the result imply a lack of robustness to other datasets or simply that thousands of errors in the MID data are skewing estimates of the variable? It would be up to the researcher to determine which is true. CoW's own estimates suggests multiple thousands of errors in their data, so it would be impossible to determine without a comprehensive review whether *any* differences between estimates are due to coding rule or errors in their data.

References

Ghosn, Faten, Glenn Palmer and Stuart A Bremer. 2004. "The MID3 data set, 1993–2001: Procedures, coding rules, and description." *Conflict management and peace science* 21(2):133–154.

Gibler, Douglas M and Steven V Miller. 2023. "The Militarized Interstate Confrontation Dataset, 1816-2014." *Journal of Conflict Resolution* p. 00220027221104704.

Gibler, Douglas M. and Steven V. Miller. N.d. "Our Perspective: Thousands of Errors Remain in the MID Data." Posted response to CoW authors' rejoinder in ISQ, available at .

Gibler, Douglas M, Steven V Miller and Erin K Little. 2016. "An analysis of the militarized interstate dispute (MID) dataset, 1816–2001." *International Studies Quarterly* 60(4):719–730.

Gibler, Douglas M, Steven V Miller and Erin K Little. 2020. "The Importance of Correct Measurement: A Response to Palmer, et al." *International Studies Quarterly* .

Gochman, Charles S and Zeev Maoz. 1984. "Militarized interstate disputes, 1816-1976: Procedures, patterns, and insights." *Journal of Conflict Resolution* 28(4):585–616.

Jones, Daniel M, Stuart A Bremer and J David Singer. 1996. "Militarized interstate disputes, 1816–1992: Rationale, coding rules, and empirical patterns." *Conflict Management and Peace Science* 15(2):163–213.

Palmer, Glenn, Vito d'Orazio, Michael Kenwick and Matthew Lane. 2015. "The MID4 dataset, 2002–2010: Procedures, coding rules and description." *Conflict Management and Peace Science* 32(2):222–242.

Sarkees, Meredith Reid and Frank Wayman. 2010. *Resort to war: 1816-2007. Correlates of War*. Washington DC: CQ Press.