Appendix to

The Militarized Interstate Dispute Dataset: Putting Things in Perspective

By

Glenn Palmer
Vito D'Orazio,
Michael R. Kenwick
Roseanne W. McManus

Table of Contents

PART 1: REPLICATIONS

Summary of the findings across different replication studies by GML (2016), PDKM (2020), and GML (2020): Assessing divergence and validity

We begin by providing a general overview of the replication findings to date in GML (2016), PDKM (2020), and GML (2020). A common goal in all of these analyses has been to determine whether the choice between using the COW MID data or GML-modified MID data leads to significantly different findings about empirical relationships that are of interest to conflict scholars.

Two questions are central to these replication studies. First, are there *systematic* differences among the findings produced by each dataset? Systematic differences in the data might suggest that there were systematic differences in the application of the coding rules across historical cases. In contrast, infrequent or inconsistent differences would suggest that the contrast between the datasets may be driven by measurement error without any systematic bias. Second, does one dataset produce results that better accord with established theory? If this is the case, this would be evidence that one dataset has better predictive validity than the other.

At this point, four studies have been replicated by GML (2016),[1] and two of these have been replicated twice more by PDKM (2020) and GML (2020), using altered research designs each time. Although scrutiny of the replications by both GML (2016) and PDKM (2020) revealed research design shortfalls,[2] we think that summarizing how the results have changed through the different iterations is a useful exercise.

Table 1 presents a list of all instances where a variable has changed in statistical significance – i.e., moved above or below the p=0.05 cutoff – when switching from the MID data to the GML data. This is an admittedly crude metric of comparison, as changes in significance are neither necessary nor sufficient for detecting meaningful changes in coefficient values (Gelman and Stern 2006). Often, these knife-edge tests exaggerate the scale of differences because coefficients sometimes fall narrowly to one side of the p=0.05 cutoff. Nevertheless, this provides an easily transferable metric to apply across different regressions and studies.

Turning to the question of whether there are systematic differences among the findings, we note that there are relatively few instances where the GML data consistently produce different results for a particular variable, either across the replications in GML (2016), PDKM (2020), and GML (2020), or across models within any one of these replications. The biggest exceptions are issue type and territory. Within the GML (2020) study only, territorial disputes are more weakly associated with escalation when using the GML data across several models in the Braithwaite and Lemke (2011) replication (see Figure 1 in GML's (2020) appendix). Similarly, the indicator for disputes over regime type is more weakly associated with reciprocation in some models in the Weeks replication. That disagreement exists on the issue type indicators is perhaps unsurprising,

[1] We are counting GML's (2016) "dangerous dyads" analysis as a replication, even though they do not aim to exactly replicate prior work.

[2] We did not closely scrutinize GML's (2020) research design or review their replication files.

since the theoretical justification and coding rules for issue type are less well developed compared to other coding criteria (see Diehl 2013). In addition, monarchies are more strongly associated with reciprocation when using the GML data across all model specifications in the Weeks replication in GML (2020), but these regimes only comprise about 2 percent of the total data, so volatility in results is perhaps to be expected. Overall, the general patterns suggest the scale of disagreement has narrowed since the original GML (2016) analysis, at least for the Weeks replication, and the disagreement appears be un-systematic.[3]

When disagreements do emerge, the next question is what this implies about the validity of the MID versus GML data. The difficulty, of course, is that one cannot simultaneously test a theory and validate a measure in a single analysis. Instead, one must *either* assume that a theoretical relationship is true and evaluate a measure's ability to reproduce it (i.e., predictive validity) *or* assume one measure is more valid and then use it to test (or re-test) theory. If the goal is to compare the validity of the MID and GML data, then the former approach is necessary. This task is made difficult because the underlying theoretical relationships must be evaluated on their *deductive* merits alone rather than their previously uncovered empirical support, since the latter was obtained using the MID data in the first place.

With this in mind, we first re-evaluate Braithwaite and Lemke's (2011) finding that there are few consistent predictors of escalation, with the exception of territorial conflict. In their original analysis, GML (2016) found that the effect of democracy was attenuated and concluded that their data brought empirics closer in line with existing theory, or lack thereof. In PDKM (2020) and GML (2020) however, the relationship between democracy is similarly insignificant across datasets, so this suggests little from a comparative validation standpoint. By contrast, while Braithwaite and Lemke (2011), GML (2016), and PDKM (2020) all found territory was consistently associated with escalation of conflict across both datasets, this relationship is attenuated using the GML data in GML (2020). If one assumes the GML data are superior to the MID data, this is problematic for theories of territory and interstate conflict. If, however, one accepts these theories on their deductive merits, it suggests superior predictive validity for the MID data compared to the GML data.

Turning to Weeks, GML (2020) note that they maintain their original finding that there is no relationship between regime type and reciprocation. While this is true, the result is not driven by differences in the MID and GML data, which perform very similarly across all major regime types, save for monarchies, which comprise a very small proportion of observations and for which theory provides little guidance from a predictive validity standpoint.

We therefore do not believe that the replication analysis performed to date suggests that either version of the dataset is necessarily more valid than the other. It is, of course, possible that future research may be able to shed more light on this question, using different designs and emphasizing different variables of interest in the MID data.

[3] In some cases, a variable's significance has changed consistently across replications, but a visual examination of coefficients shows that the estimates themselves are not substantively different from each other. This is especially true for power preponderance in the Braithwaite and Lemke replication, and major power status in the Weeks replication.

Table 1: Coefficients that Gained or Lost Statistical Significance when Changing from the MID Data to the GML Data

	Braithwaite and Lemke (2011) DV: Escalation 6 Models 84 Coefficients	**Weeks (2008)** DV: Reciprocation 4 Models 64 Coefficients
GML (2016)	• Defense pact becomes positive and significant in stage 2 (M3) • Joint satisfaction loses significance in stage 2 (M5) • Power preponderance loses significance stage 2 (M5) • Rho becomes positive and significant (M2) • Rho loses significance (M6)	• Democracy loses significance (M1) • Personalism loses significance (M2, M3) • Nondemocratic interregna loses significance (M2, M3) • New democracy loses significance (M2) • Major-minor loses significance (M2) • Alliance portfolio similarity loses significance (M3) • Territory become positive and significant (M1, M2)
PDKM (2020)	• Power preponderance loses significance in stage 1 (M1, M2, M4, M5, M6) • Joint democracy becomes positive and significant in stage 2 (M1, M5) • Joint satisfaction loses significance in stage 2 (M2) • Territorial dispute becomes positive and significant in stage 2 (M5) • Power preponderance becomes positive and significant in stage 2 (M6)	• New democracy becomes positive and significant (M2) • Military regime becomes positive and significant (M2) • Alliance portfolio similarity becomes positive and significant (M2, M3) • Personalist loses significance (M3) • Nondemocratic interregna becomes negative and significant (M4)
GML (2020)	• Power preponderance loses significance in stage 1 (all six models) • Power preponderance gains significance in stage 2 (M5, M6) • Joint satisfaction loses significance in stage 2 (M2) • Territorial dispute loses significance in stage 2 (M1, M3, M4, M6)	• Non-dynastic monarchy becomes positive and significant (M2, M3, M4) • Dynastic monarchy becomes positive and significant (M2, M3, M4) • Regime dispute indicator becomes insignificant (M1, M2, M4) • Major-major power becomes negative and significant (M1)

Note: We count statistical significance changes using the p=0.05 cutoff. For GML (2016), we compare the GML data to the MID 3 data. For PDKM (2020) and GML (2020), we compare the GML data to the MID 4.3 data. For GML (2020), this comparison is based on visual inspection of their figures. Furthermore, for GML (2016), we compare GML's replication using their own data with GML's replication of Braithwaite and Lemke using the MID 3 data, not with Braithwaite and Lemke's original results, which neither team has been able to replicate exactly. We do not analyze the peace years coefficients for Braithwaite and Lemke because they are not reported in any of the results. We did not count the differences in the alliance and contiguity indicators from Weeks M4 because it appears as though these coefficients were reversed in the visual display reported by GML. If we find that this is not the case, we will update the table accordingly.

Braithwaite and Lemke (2011) Replication

GML (2020) suggest there are a large number of errors in our replication of Braithwaite and Lemke. We do acknowledge two errors related to treatment of missing values in GML's data and reconstruction of the territorial dispute dummy, but we disagree that these two errors drove our results. We view the additional issues that GML raise with the replication as research design choices rather than errors, and we note that all of these research design choices were carried over from the previous analyses.

It is important to note that none of the replication efforts have succeeded in replicating the exact dataset construction and results published by Braithwaite and Lemke. The primary goal in our replication was to construct the dependent variables as similarly as possible to Braithwaite and Lemke, even if this was not the way in which we would have chosen to construct them if we had designed the original analysis.

Dataset Versions: GML (2020, 4) state, "PDKM added to these errors by using inconsistent MID data versions across levels of analysis—MIDA 3.0 but MIDB 3.02." This is incorrect. We used MID 3.0 consistently throughout the Braithwaite and Lemke replication, as shown in our replication files. We did use MID 3.02 in the Weeks replication because this is what Weeks used. Additionally, GML (2020, appendix page 8) express skepticism that we could have actually used the MID 3.0 dataset because they were unaware of its existence. We cannot verify whether version 3.0 was ever published on the COW site, but it exists in the MID project archives, and this is what Braithwaite and Lemke (2011) say they used, so we used it also. This dataset is now publicly available with our ISQ replication data.

Coding Democracy: GML state that our finding regarding the effect of democracy is surprising and trace this back to an error in coding democracy in the original work by Braithwaite and Lemke (2011). We have no objection to GML (2020) recoding this variable, but our own replication approach – which was the same as GML's approach in their 2016 analysis – was not to adjust Braithwaite and Lemke's coding of any variables that are not related to MIDs.

Coding MID Onset: GML argue that we did not adequately investigate the reasons why their coding of MID onset differed from Braithwaite and Lemke's and that many of these differences were due to changes in the dataset rather than research design choice. In response, we note that we never stated that all of the differences were due to research design and never believed this was the case. In retrospect, we realize that listing the number of differences in our ISQ response without explicitly stating the number that were due to research design could have given this impression. Therefore, we admit a shortcoming in our written explanation. Ultimately though, we and GML are in agreement that some of the differences were due to GML's dataset changes and others were due to their research design (see GML's 2020 Appendix, page 8, Table 2). Our main point was not about the number of observations that changed due to research design, but rather the effect that this had on the results.

Coding Escalation: GML (2020) make three major critiques of our coding of escalation. First, they criticize the fact that we did not always account for joiner behavior when coding escalation, suggesting that this is inconsistent with our prior argument that not only bilateral disputes should

be counted. Second, they point out that we do not always treat missing fatality values consistently, sometimes treating them as escalations and other times as non-escalations. Both of these critiques misunderstand our replication strategy. In all cases when we coded escalation, we experimented with different coding methods in order to match Braithwaite and Lemke's as closely as possible. We were able to match their coding of mutual force perfectly, and we achieved reasonably good matches for the other escalation variables. We prioritized matching Braithwaite and Lemke over being philosophically consistent or even objectively correct in our coding. Given this goal, we believe we coded the variables as correctly as possible.

GML's third critique is that we failed to convert some of their -9 codings to missing values because we used the mvdecode command before the destring command with their dataset. We agree that we did make this mistake. However, the fact that we only made this mistake with the GML data and not the MID data (which did not need to be destringed) would create bias *away* from finding the similar results between the MID and GML data that we actually found.

Coding Peace Years: GML (2020) criticize our analysis for not reconstructing the peace years variables with each of the different datasets. We deliberately chose not to reconstruct the peace years variable because GML (2016) did not do so, and we were trying to replicate their analysis as closely as possible except for the changes that we highlighted.

Territorial Dispute Coding: GML (2020) also criticize our analysis for not reconstructing the territorial dispute variable. We acknowledge this as our second error. We agree that we should have reconstructed this variable, since GML (2016) use a reconstructed version of it in the second-stage of their regressions (although they use Braithwaite and Lemke's original variable in the first stage).

Impact: While we view some of GML's assertions above as having merit and others as lacking merit, we believe that all of these considerations have minimal impact on the substantive findings. After GML (2020) make adjustments to our analysis, they show that their changes to the MID dataset cause 13 out of 84 coefficients to move above or below the statistical significance threshold, which is only three more significance changes than we found in PDKM (2020). Moreover, as we discussed in Section 1, both datasets now suggest an insignificant relationship between democracy and escalation, which is contrary to GML's (2016) original contention that their data better captured a theoretically plausible null relationship. GML's (2020) data does attenuate the relationship between territorial disputes and escalation, but this does not suggest superior predictive validity, since it seemingly contradicts theories suggesting a strong positive relationship (Vasquez 1993; Senese and Vasquez 2008). Ultimately, both our analysis in PDKM (2020) and GML's (2020) analysis support a similar conclusion that differences in results between the two datasets are not very common.

Weeks (2008) Replication

GML (2020) make three claims with regard to our analysis of their previous replication of Weeks (2008). First, they deny our claim that their original analysis (GML 2016) misattributed dispute-level reciprocation to dyadic-level data. Second, they claim that our analysis (PDKM 2020) dropped a subset of cases that created potential bias against finding differences among the datasets. Third, they claim that this mistake drove our core finding that there were few differences in the results across datasets. Only the second claim is correct; we failed to account for the possibility of a selection effect based on observations that were dropped in our replication. However, this selection effect was not enough to drive our finding that differences between the MID and GML data seldom affect results. This persists even in GML's (2020) update of their original replication (GML 2020, Appendix, p. 14).

Turning to the first claim in more detail, GML (2016) did not release replication files showing how their data were constructed, but in the replication files for our original response (PDKM 2020), we showed that merging the dispute-level reciprocation variable from GML's dataset into the dyadic Weeks (2008) data produced a variable that matches GML's (2016) reciprocation variable perfectly in 2,269 of 2,270 observations. There is a single anomalous case (MID 2802) where our reproduction indicates reciprocation, while GML's indicates no reciprocation, as we pointed out in our original replication files. After adjusting this case and including one additional MID observation that GML indicate should be dropped but was contained in their replication analysis, we are able to perfectly replicate GML's (2016) results using the dispute-level reciprocation variable that we reconstructed from their data, demonstrating that they did in fact merge the dispute-level variable into Weeks' dyadic data in their original analysis. A case-by-case investigation of individual observations further confirmed the existence of observations where Recip was coded as one, even though there was no reciprocation within the dyad, according to GML's data.

Regarding the anomalous case of MID 2802, we confirmed that GML coded it incorrectly according to their own dataset – both their MIDA and MIDB data indicate that the dispute is reciprocated. We also found that this anomalous case is consequential, at least for knife-edge significance testing. GML (2016) find that, contrary to Weeks, personalism is no longer significantly associated with reciprocation in any specification, which they attribute to their updates to the MID data. However, if MID 2802 is coded correctly according to their data, personalism is again significant using the GML (2016) data at the $p<0.1$ level in Weeks' Model 2 and at the $p<0.05$ level in Weeks' Model 3.[4]

Turning to our own analysis, GML (2020) is correct to say that when we re-constructed GML's (2016) reciprocation variable with a correction for the dyadic nature of the dispute data, we did indeed drop 159 observations in which Side A and Side B were switched or one of the participants changed.[5] We will first explain why we proceeded as we did and then discuss potential problems

[4] These p-values were originally 0.102 and 0.057 respectively in GML (2016). GML do not remark on significance at the $p<0.10$ level, but neither does Weeks (2008).

[5] Contrary to GML's assertion, no observations where the start year changed were dropped because we did not merge on the start year.

with this approach. In their original replication of Weeks, GML (2016) used Weeks' replication dataset and merged their dispute-level reciprocation variable into it. Seeking to replicate GML as closely as possible, except for making the reciprocation variable dyadic, we also merged our corrected reciprocation variable into Weeks' original dataset. It did not occur to us to match cases in which Sides A and B were switched or in which one of the participants changed entirely because all of the country-level independent variables in the Weeks' data would now be incorrect for these observations, having been coded for the previous Side A and B. Thus, GML's (2016) inclusion of these observations in the regression, without adjusting the independent variables, could only create bias away from finding the true result.

While we think that GML's (2016) inclusion of these observations was problematic, we do acknowledge GML's (2020) point that excluding them was also potentially problematic. Since these observations all contain changes between the datasets, simply ignoring their existence could create bias against finding a difference between the datasets. That said, we do not believe that this bias was the driving force behind our primary conclusion that changes in the MID data do not alter the substantive findings of Weeks because GML (2020) now show a similar finding.

Ultimately, we think that GML's (2020) decision to entirely reconstruct the Weeks dataset from scratch is a good one,[6] since it allows the dyadic reciprocation variable to be merged in along with correct independent variables. However, this and other changes in design that GML (2020) make have important repercussions. GML (2020) are no longer able to reproduce the core findings from Weeks (2008) using any current or previous version of the MID or GML data. Meanwhile, our core finding continues to hold: the differences between the MID and GML data rarely produce meaningfully different results. In the updated GML (2020) analysis, there are 10 instances (out of 64 possibilities) where a variable moves above or below the significance threshold as a result of switching to the GML data, and 6 of these pertain to dynastic and non-dynastic monarchies, which comprise approximately 1.4 and 0.5 percent of the data, respectively.[7] GML (2020) claim that their original findings are supported because Weeks' substantive findings are invalidated, but this is no longer due to changes they made to the MID data, as they originally claimed. Although GML's (2020) findings go against Weeks by showing no significant effect of personalist regimes, this is now true for both their dataset and the MID dataset.

Gibler and Miller (2013) Replication

GML (2020) correctly note that we did not provide justification for our decision not analyze their replication of Gibler and Miller (2013). The reason was that they indicated the findings they report were due partly to their decision to drop missing start dates when using the MID data and infer them with the GML data. Because the analysis conflated changes to design and changes to the MID data, we did not scrutinize the results further.

[6] Although we did not check their methodology for doing this.
[7] Percentages are based on the original Weeks (2008) data, but are unlikely to change in any subsequent analysis.

PART 2: CODING

Events that Cannot be Found

A core issue in our disagreements with GML relates to the issue of events in the original MID data that GML could not find. Our reasons for retaining the 19 entire disputes that cannot be found were already explained in our previous response. Similar reasoning contributes to our reluctance to accept GML's other recommendations without review. In cases where GML suggest changes to dates, action types, and even participants, we want to attempt to verify whether GML is actually capturing *all* of the incidents within the original MID. Finding documentary evidence of one incident does not necessarily mean that other incidents did not occur, and if other incidents occurred, then it is possible that the original MID coding is correct.

Disagreements in Applying the Coding Rules

GML (2020) highlight several instances where they disagree with our decision not to accept their recommended changes to the classification of MIDs. GML's discussion magnifies the extent of disagreement since they select on the remaining subset of cases we have mutually reviewed and disagree on, and further on cases where they find their arguments against our classifications most compelling. The appendix to our original study already contains case-specific explanations for our decisions and we will not resurrect these debates here; we simply take note of some general disagreements with passing references to illustrative examples.

Private Citizens as Targets: One somewhat consistent area of disagreement pertains to whether a militarized action targets another state entity or a private citizen. The coding rules state that actions against private citizens outside their home territory are generally not coded as incidents with exceptions including (but not limited to) seizures in disputed territory, attacks on international shipping, and pursuit of rebel forces, and here only when the "targeted" state responds militarily. Nevertheless, knowing whether a particular action targets private citizens or another state can be context dependent. This is frequently the case with shows of force, defined as "public demonstrations by a state of its military forces intended to intimidate another state not involving military operations." As the coding manual states, these include non-routine naval patrols and the intentional violation of territorial waters or airspace. Often we observe countries violating each other's (or disputed) airspace and harassing private crafts. Whether such actions constitute MIDs can be context dependent, though when two states share rivalrous relations (in a colloquial sense), we find it more reasonable to infer that such maneuvers are state-to-state signals of resolve. Put differently, Russia scrambling its jets near Alaska is more likely a show of force against the United States than it is an attempt to intimidate a private airline corporation or the individuals on those flights.

Propaganda: GML (2020) contend that the MID team consistently codes propaganda, but this is not the case; we never knowingly incorporate information into the data that we know is false. That said, descriptions of interstate conflict are rife with competing accounts and inconsistent information, and we are cautious to label news sources as propaganda unless we have significant evidence that this is so. For example, GML contend that Libya manufactured an incident in which U.S. jets allegedly buzzed their aircraft. GML quote U.S. military personnel who they believe

convincingly deny the claim. Our team was not sufficiently convinced by this denial. The U.S., like any country, often has strategic incentive to deny its actions.

We disagree with GML that this is an obvious misclassification and believe it is one of many difficult cases on which reasonable and judicious coders can respectfully disagree. We do not, and never have claimed to have measured all disputes perfectly in the presence of incomplete information, but we do apply the coding rules as judiciously as possible in these cases. Still, we do not believe GML have produced evidence that we are systematically more likely to code propaganda, nor that we are "incorrect" in our classifications along these dimensions.

Blockades: GML (2020) say "we cannot understand why PDKM would argue that a state closing its border to traffic from another country constitutes a blockade (3 cases). That makes no sense, especially since CoW coding rules describe blockades well." The coding manual defines blockades as follows:

> ***Blockade*** — the use of military forces by one state to partially or completely seal off the territory of another state to prevent the entry or exit of goods or personnel. Stopping or inspecting ships or vehicles or the confiscation of goods in transit is evidence of the erection of a blockade. A formal declaration is not required.

Thus, closing border crossings and halting border traffic would constitute a blockade so long as it is done by regular military forces. We found that GML would sometimes recommend that a dispute should be eliminated because border crossings were closed but would not mention whether military or non-military forces were used, so we did not find the recommendation actionable.

Reciprocation, Clashes and Attacks: GML (2020) argue against the use of dispute reciprocation as a substantively meaningful variable for testing audience costs based on their reading of the MID coding rules. They quote the original definition of an attack as:

> Clash: outbreak of military hostilities between regular armed forces of two or more system members, in which the initiator may or may not be clearly identified.
> Raid:[8] use of regular armed forces of a state to fire upon the armed forces, population, or territory of another state. Within this incident type, the initiator can be clearly identified and its action is not sanctioned by the target.

GML then argue, "The difference between coding a clash versus an attack is whether the initiator could be identified" (GML 2020, Appendix, p. 15).

This does not square with our application of the coding rules. To our knowledge, MIDs have never been coded using such a decision rule. The updated language in the incident coding manual used since MID 3 makes the distinction clear: clashes are reciprocated, whereas attacks are not.

> ***Clash*** — the outbreak of sustained military hostilities between the regular armed forces of two or more states. This differs from an Attack, a unilateral action, in that a Clash is reciprocal in nature. The initiator (the Actor side) may or may not be clearly identified,

[8] Raid was the original name for actions that are called attacks in the current MID data.

> and it will be assumed that the designation of Actor(s) and Target is arbitrary in this type of incident unless stated otherwise in the Notes section.

Sources of Information

GML (2020) write:

> As PDKM note, they relied on our brief summaries of these cases or encyclopedic treatments of conflicts, not primary sources. They also write that they relied on the book of MID narratives we produced (Gibler, 2018), but it actually contains no discussion of these dropped cases.

This statement is not correct. In our review we read both primary and secondary documents extensively. Over 150 of these are listed on pages 38-47 of our original appendix. Many, though not all, of these were originally reported by GML, and we credited them for this advancement. Similarly, we referred to Gibler (2018) as a significant step forward in improving the measurement process, but never implied that we read the narratives contained therein to make judgements about dropped MIDs.

PART 3: COUNT OF DIFFERENCES BETWEEN THE DATASETS

GML's (2020) Tables 1 and 2 exaggerate the number of differences between the MID and GML data both (1) by counting changes to variables that are not independently coded but rather derived from other variables and (2) by counting several of the exact same changes multiple times. Our Tables 2 and 3 below show how the number of differences changes after addressing these issues.

Table 2: Amended Version of GML's (2020) Table 1, Counting Dispute-Level Differences

Difference Type	Count of differences with MID 4.3 from GML (2020) Table 1	Eliminate double counts and variables not coded independently	Reason omitted
Different fatalpre	217	217	
Different hiact	445	445	
Different hiact AND hostlev	140	omit	Double count
Different hostlev	140	omit	Derived from hiact
Different numa	73	73	
Different numa AND numb	36	omit	Double count
Different numb	84	84	
Different duration (max)	1,174	omit	Derived from start and end dates
Different dur (min AND max)	1,145	omit	Double count
Different duration (min)	1,149	omit	Derived from start and end dates
Different end dates	1,101	1,101	
Different end day	905	omit	Double count with end date
Different end month	484	omit	Double count with end date
Different end year	126	omit	Double count with end date
Different fatality	258	258	
Different fatality AND fatalpre	175	omit	Double count
Different reciprocation	250	250	
Different start AND end dates	685	omit	Double count
Different start AND end year	19	omit	Double count
Different start dates	978	978	
Different start day	795	omit	Double count with start date
Different start month	324	omit	Double count with start date
Different start year	55	omit	Double count with start date
MIDs dropped or merged	128	128	
Total changes	10,886	3,534	
Total changes reviewed	128	128	
Total changes un-reviewed	10,758	3,406	

Note: We follow GML (2020) in not counting changes to outcome and settlement. We reviewed 269 MID drop recommendations and 75 merge recommendations, so the 128 "total changes reviewed" refers to the total that we still disagree about after review.

Table 3: Amended Version of GML's (2020) Table 2, Counting Participant-Level Differences

Difference Type	Count of differences with MID 4.3 from GML (2020) Table 2	Eliminate double counts and variables not coded independently	Reason omitted
Different fatalpre	245	245	
Different hiact	962	962	
Different hiact AND hostlev	510	omit	Double count
Different hostlev	511	omit	Derived from hiact
Different orig	48	48	
Different sidea	400	400	
Different end dates	2,768	2,768	
Different end day	2,065	omit	Double count with end date
Different end month	1,195	omit	Double count with end date
Different end year	383	omit	Double count with end date
Different fatality	333	333	
Different fatality AND fatalpre	244	omit	Double count
Different start AND end dates	1,763	omit	Double count
Different start AND end year	95	omit	Double count
Different start dates	2,418	2,418	
Different start day	1,794	omit	Double count with start date
Different start month	786	omit	Double count with start date
Different start year	185	omit	Double count with start date
Missed or incorrectly identified participants	118	118	
Participants dropped or merged	489	489	
Total changes	17,312	7,781	
Total changes reviewed	0	0	
Total changes un-reviewed	17,312	7,781	

Note: GML (2020) omit changes relating to revision from their table, noting in the appendix that they are still working on a fuller list of these. We follow their lead.

Tables 2 and 3 above show that the count of differences at the dispute level is reduced by around two thirds, and the count of differences at the participant level is reduced by more than half, after making these adjustments. The number of substantive differences in our count is still slightly exaggerated because we count Fatality as being coded independently from Fatalpre, although this is true only in cases where the exact fatalities are unknown but a rough level can be estimated.

It should also be noted that the MIDA dispute-level dataset is entirely derived from the MIDB participant-level dataset. Therefore, while counting differences at both levels provides some informational value, the total differences that remain to be reviewed is **not** the sum of differences at both levels, but rather merely the total participant-level changes.

Finally, it is important to put the number of differences between the datasets in context with the number of substantive coding decisions with potential for disagreement. In calculating this, we

note that potential for disagreement exists over **both** the existence of each observation in the dataset (i.e., whether it is truly a MID or a MID participant) **and** over each of the individual variable values within each observation. As before, however, we do not wish to double-count variables that are merely derivations of other variables. Therefore, our formula for calculating substantive coding decisions with potential for disagreement is:

*Substantive coding decisions = number of observations + (number of observations * number of non-omitted variables in Table 2 or 3 above)*

We use this formula to make the calculations shown in Table 4, based on the number of observations in the MID 4.3 data. Comparing the adjusted total differences in Tables 2 and 3 with the number of substantive coding decisions in Table 4 suggests a disagreement rate of approximately 16 percent at the dispute level and 17 percent at the participant level. While this is a non-trivial disagreement rate, it is important to keep in mind that not all alternations to the dataset will affect all analyses equally. This is why much of our own analysis focuses on the substantive impact of the changes in replication studies.

Table 4: Calculating the Number of Substantive Coding Decisions in MID 4.3

	MIDA (Dispute Level)	MIDB (Participant Level)
Number of non-omitted variables from tables above	8	7
Number of observations	2,315	5,558
Number of substantive codings with potential for disagreement	**20,835**	**44,464**