Updating the Militarized Interstate Dispute Data: A Response to Gibler, Miller, and Little

GLENN PALMER
The Pennsylvania State University

VITO D'ORAZIO
The University of Texas at Dallas

MICHAEL R. KENWICK
The University of Pennsylvania

ROSEANNE W. MCMANUS
The Pennsylvania State University

ABSTRACT

In a recent article, Gibler, Miller & Little (2016) (GML) conduct an extensive review of the Militarized Interstate Dispute (MID) data between the years 1816 and 2001, highlighting possible inaccuracies and recommending a substantial number of changes to the data. They contend that, in several instances, analyses with their revised data lead to substantively different inferences. Here, we review GML's MID drop and merge recommendations and reevaluate the substantive impact of their changes. We are in agreement with about 76 percent of the recommended drops and merges. However, we find that some of the purported overturned findings in GML's replications are not due to their data, but rather to the strategies they employ for replication. We re-examine these findings and conclude that the remaining differences in inference stemming from the variations in the MID data are rare and modest in scope.

We are grateful to a number of research assistants who worked to both verify and describe the changes to the MID data. These include Benjamin Ostick, Matthew Lane, Dennis Okyere, Sarah Parrott, and Zhanna Terechshenko. We are particularly grateful to Kim Tran, who worked diligently to compose many of the narratives contained in our appendix.

The Militarized Interstate Dispute (MID) data are among the most widely used and analyzed datasets in the study of international conflict and international relations more broadly. Thorough evaluation and validation of these data are critical tasks for advancing the scientific study of war and peace. Gibler, Miller & Little (2016) (hereafter GML) conduct such an evaluation and recommend removing 269 disputes from the MID data, merging 75 disputes together with other existing disputes, and making alternations to over 1,000 additional disputes.[1] They contend that these changes, in many instances, lead to substantively consequential differences in empirical analyses.

In this article, we review GML's MID drop and merge recommendations and reexamine GML's claim that their revisions lead to new inferences. We focus on the drop and merge recommendations because these decisions are most likely to impact empirical analyses. To briefly summarize, of the 269 MIDs that GML recommend dropping, we are in agreement for 200 cases (74 percent). Among the 75 cases where GML recommend merging disparate MIDs together, we agree with 62 (83 percent) of these merges.

Next, we re-replicate two of the replication studies in GML and compare the findings with the original MID data, the GML data, and the revised MID 4.3 data, which incorporate the GML recommendations that we accepted. While GML state that their data produce substantively different results, we find that some of the most consequential changes that GML identified are driven not by differences among the datasets, but rather by different methods of constructing the dependent variables. When coding the dependent variables in a manner consistent with the original analyses, we find that which dataset is used makes little substantive difference in the findings. This suggests that despite some "noise" in the MID data due to potentially different interpretations and application of the coding rules, the data nonetheless allow us to make meaningful and consistent inferences.

This exercise highlights the importance of measurement validation in social science research. Like many concepts of interest in international relations, the MID (and the notion of a "dispute" in general) is a complex social construct. There is often considerable uncertainty regarding interstate interactions and

[1] In the text of their article, GML state that they recommend 251 drops and 72 merges (while their Tables 1 and 2 list 245 drops), but their associated appendix lists 250 drops and 75 merges. We report these latter numbers. We also add the 19 MIDs on their "could not find" list to the 250 drop recommendations in this summary, since GML ultimately recommend dropping them as well. It should also be noted that GML have updated their recommendations since their article was published, but we confine ourselves to addressing GML's published recommendations for purposes of this response.

coders must attempt to adjudicate between competing claims and reports from various agencies to resolve ambiguities relating to the precise nature, intent, and target of particular actions. Assessing the importance of these coding decisions, especially in terms of their impact on statistical inferences, is an important exercise. Moving forward, we are dedicated to improving the transparency and reproducibility of the MID data.

Summary of Changes

The changes recommended by GML can be placed into four categories: (1) 19 MIDs that GML recommend dropping because they could not recover any historical evidence of a militarized incident; (2) 250 MIDs for which GML recover historical information, but recommend dropping because they argue that the actions involved do not meet the criteria outlined by the MID coding rules; (3) 75 MIDs that GML recommend merging into other existing disputes; and (4) over one thousand MIDs that GML recommend keeping with alterations to the original coding. Evaluating each of these categories poses several challenges, particularly when they relate to data from early iterations of the MID project, which began in the 1960s when norms of data replication in political science were less well developed and the technology for data maintenance was more primitive than in the present. Although documentation is available that broadly lists the sources used to compile information about each militarized dispute, it is often challenging to locate the specific news article or book passages that informed the original coding decision. We are therefore sympathetic with the difficulty of the process undertaken by GML and believe their review and the documentation in Gibler (2018), represent an important step forward. Nevertheless, in the cases we reviewed, we disagree with about 24 percent of their recommended changes.[2]

A discussion of the process undertaken to evaluate GML's revisions is described in the online appendix. Interested readers should also consult the appendix for a discussion and complete list of cases where we agreed and disagreed with GML's recommendations to drop, merge, and revise.

[2] Total reviewed disputes include the 19 recommended drops for which there is no documentation, the 250 recommended drops for which there is documentation, and the 75 recommended merges. We agree with 262 (76 percent) of these 344 revisions.

MIDs that GML Could Not Find

GML recommend dropping 19 observations for which historical information could not be recovered. Because GML were unable to locate information on these events, they assume that the original coders either miscoded some aspect of the event or mistakenly identified a non-militarized incident as a MID. Determining how to treat these observations is difficult given the minor nature of many militarized incidents, which often entail actions such as border fortifications, violations, and shows of force. Low-level interactions such as these are not always well documented in historical texts. This problem is most acute when assessing MIDs from the nineteenth and early twentieth centuries. Even the internet has limited utility for finding historical information about small events.

The following anecdote highlights this difficulty. During a 2014 workshop with GML, their team initially identified 33 observations in this category. During this meeting, information was recovered on ten of these disputes, six of which GML ultimately decided to keep in the data. Since that time, the number of undocumented MIDs has been further reduced to 19. We believe this reduction is suggestive of how difficult it can be to recover information on some historical MIDs, even when these events did, in fact, take place.

Thus, while we cannot be certain that the original coding is correct, it is highly likely that at least some of these 19 disputes reflect true historical events. Therefore, we argue that excluding all 19 has a greater potential to create bias than including all 19. In sum, in the absence of information, we err on the side of inclusion rather than exclusion and retain these 19 MIDs.

MIDs with Documentation that GML Recommend Dropping

There are an additional 250 observations that GML have recovered information on, but believe should be removed from the MID data due to misclassification of non-militarized events by the original coders. We paid particular attention to examining these cases because they are the most likely to impact empirical analyses that use the MID data. Of the original 250 cases, we are in agreement with GML that 200 (80 percent) should be dropped. We have dropped these MIDs from the latest version of the data (MID 4.3). We have determined that the remaining 50 cases should remain in the MID data. In the online appendix, we detail each of these cases and provide a rationale for our decision.

Our decisions to retain these cases can be placed into three categories: (1) cases in which we identified relevant information beyond that which GML used, (2) cases in which we disagree with GML's interpretation of events, and (3) cases in which we disagree with GML about the application of the MID coding rules. An example of the first category is MID #4023, an alert by China targeted at Taiwan in January 1993. GML contend that neither the MID-listed sources, nor any other newspaper sources, describe such an alert. However, we found a Hong Kong newspaper article stating that Chinese armed forces had been placed on combat readiness in response to events in Taiwan and giving details about which units were on alert (Hui-wen 1993). Therefore, we retained this MID.

An example of the second category is MID #1157, which occurred between Ecuador and Peru in August–September 1955. The highest action is coded as a border violation by each state. GML retrieved a *New York Times* article from September 14, 1955, which states that a four-nation committee of guarantors conducted reconnaissance and found no evidence of Peruvian troop concentrations alleged by Ecuador. GML therefore argue there is no MID. We decided to keep this MID for several reasons: First, the NYT article only pertains to one possible incident, taking place before the end date of the dispute. Second, the reconnaissance missions did not prove Peruvian troops were never at the border, only that there was no evidence of this at the time of the missions. Third, while the evidence of this troop movement is ambiguous, there is a target protest, which justifies the inclusion of such cases according to the MID coding rules. Fourth, the highest action recorded in this dispute is not actually a border fortification, but rather a border violation, suggesting the likelihood of an additional militarized incident that is part of this dispute. Finally, the NYT references Peru's arrest of four Ecuadorian soldiers on August 10, which also suggests the possibility of an additional militarized incident (*New York Times* 1955).

The third category of cases consists of instances where we disagree over the application of the coding rules to a particular event. MID #508 is one such example, for which GML state: "France said that it would not interfere in the war between Austria and Italy as long as Austria did not advance farther than Milan. France offered to arbitrate in Austria's favor if Austria agreed to this. There was no militarized incident." We believe, however, that France's statement constituted a threat, signaling that France would interfere militarily in an ongoing conflict if Austria advanced further than Milan.

The cases discussed above are atypical because, as previously stated, we agree with GML's recommendations in a majority of cases. Nevertheless, we are particularly cautious about eliminating data points without clear justification. As noted above, while the original data surely contain errors, we err on the side of inclusion because we are concerned that the elimination of cases might increase, rather than reduce, biases.

MIDs GML Recommend Merging

Militarized incidents are aggregated together into a single MID if they are contested over the same issue, take place in the same geographic area, and are systematically linked to one another. GML identify 75 MIDs that they assert contain the same actors, actions, and issue-type as already-existing MID observations and should therefore be merged. We accept 62 (83 percent) of these recommendations and incorporate them into the latest version of the MID data (MID 4.3). We did not accept the remaining 13 cases because they were predicated on other changes GML made in the MID data, including start and end date changes, that we concluded we could not accept.

MIDs GML Recommend Changing

GML identify 229 cases where they recommend changes to the dispute years, participants, or fatality levels of existing disputes. They also identify 1,009 disputes where changes are recommended for minor fields such as the start and end days. We conducted a preliminary evaluation of a sample of these changes and found that we agreed with only a minority of the recommendations and either disagreed or were unable to find sufficient evidence to support the changes in the remaining cases. However, a full review of these recommendations is beyond the scope of this article. It would require that the events coded in the original data be identified, the events in the GML data be verified, and any inconsistencies be resolved. Our goals in this article are to review the recommendations with the highest probability of affecting the results of empirical analyses (the actual elimination of disputes) and to reevaluate GML's claim that the revised data lead to substantively different conclusions.

Empirical Analysis

To assess whether the changes that GML recommend are likely to affect typical empirical analyses, we reassess GML's replications of Braithwaite & Lemke (2011) and Weeks (2008) using (1) the MID 3 data used by the original authors; (2) the GML data, which incorporate all of GML's recommendations; and (3) the MID 4.3 data, the latest version of the MID data, in which the GML changes that we evaluated and agreed with are incorporated. We find minimal substantive differences across the three datasets in each replication. We also find that some of the overturned findings that GML report are due to GML using different methods to construct their dependent variables, rather than to differences in the datasets themselves.

Braithwaite and Lemke Replication

Braithwaite & Lemke (2011) use the MID data to conduct a two-stage analysis of dispute onset and escalation. Escalation is measured using a variety of MID features, including whether a MID: was reciprocated; featured use of military force; featured use of force by both sides; resulted in fatalities; resulted in over 250 fatalities; and resulted in war. As GML note, they were unable to exactly replicate Braithwaite and Lemke's original results.

The primary difference between the original analysis and GML's analysis relates to how the researchers code dispute onset. Braithwaite and Lemke's coding of dispute onset counts all MIDs that were bilateral on day one. In contrast, GML's coding of dispute onset counts only MIDs that remain bilateral throughout their entire history. Although perhaps justifiable from a theoretical perspective, GML's different strategy for coding the dependent variable is problematic for replication. Because Braithwaite and Lemke estimate a two-stage model, GML's more restrictive coding of the first-stage dependent variable causes the second stage to be estimated on a more limited sample. For example, Braitwaite and Lemke's original sample included 453 MIDs with mutual use of force, 221 MIDs with over 250 fatalities, and 59 MIDs that escalated to war. In contrast, GML's replication includes only 399 MIDs with mutual use of force, 36 MIDs with over 250 fatalities, and 24 wars.

We re-replicate Braithwaite and Lemke's study, following the original authors' method of coding dispute onset and using (1) the MID 3.0 data originally used by Braithwaite and Lemke, (2) the GML data,[3] and (3) the MID 4.3 data. Although we are also unable to exactly reproduce the results reported by Braithwaite and Lemke using the MID 3.0 data, we find that our replication yields estimates that are closer

[3] We use the most recent GML data, version 2.1, which contains some updates compared to the version used in the original GML article.

to the original results than GML's replication.[4] Figure 1 reports the results of the replication using each of the three datasets. In each model, the results change little across the three datasets.

GML highlight the fact that joint democracy loses significance in their replications as evidence of the quality of their data, arguing that Braithwaite and Lemke's finding that this variable is a significant positive predictor of some forms of escalation is not in keeping with theory or previous findings (GML, 724). However, the insignificance of joint democracy in GML's results is driven by their method of constructing the dependent variable and not by their data.[5] In our re-replication using Braithwaite and Lemke's method of constructing the dependent variable, we find that joint democracy is more commonly a positive and significant predictor of escalation when using GML's data (significant in four models) than when using either version of the MID data (significant in two models).

GML also emphasize that the estimate of ρ, the correlation between the error terms of the equations, changes in two regressions in their analysis and argue that this is because of disputes that are dropped in their data. Again, however, we find that these changes are due to GML's method of constructing the onset variable and not their data, as our results show that ρ never changes significance across the GML, MID 3.0, and MID 4.3 datasets when using Braithwaite and Lemke's method of coding onset.

In sum, while there are a few variables that differ in statistical significance among the three datasets, these differences do not have a great impact on scholarly knowledge about dispute escalation. Given that Braithwaite and Lemke's overarching conclusion was that most predictors of conflict escalation do not have consistent effects across model specifications, the few differences that we do observe using the GML data serve to further underscore that point.

[4] Our results using the MID 3.0 data differ from Braithwaite and Lemke's results in two notable ways: First, joint satisfaction with the status quo is no longer a significant negative predictor of reciprocation. Second, power preponderance becomes a significant positive predictor of war.

[5] Indeed, GML's Table 5 (p. 275) shows that GML fail to replicate Braithwaite and Lemke's findings about joint democracy even when using the MID 3.0 data.

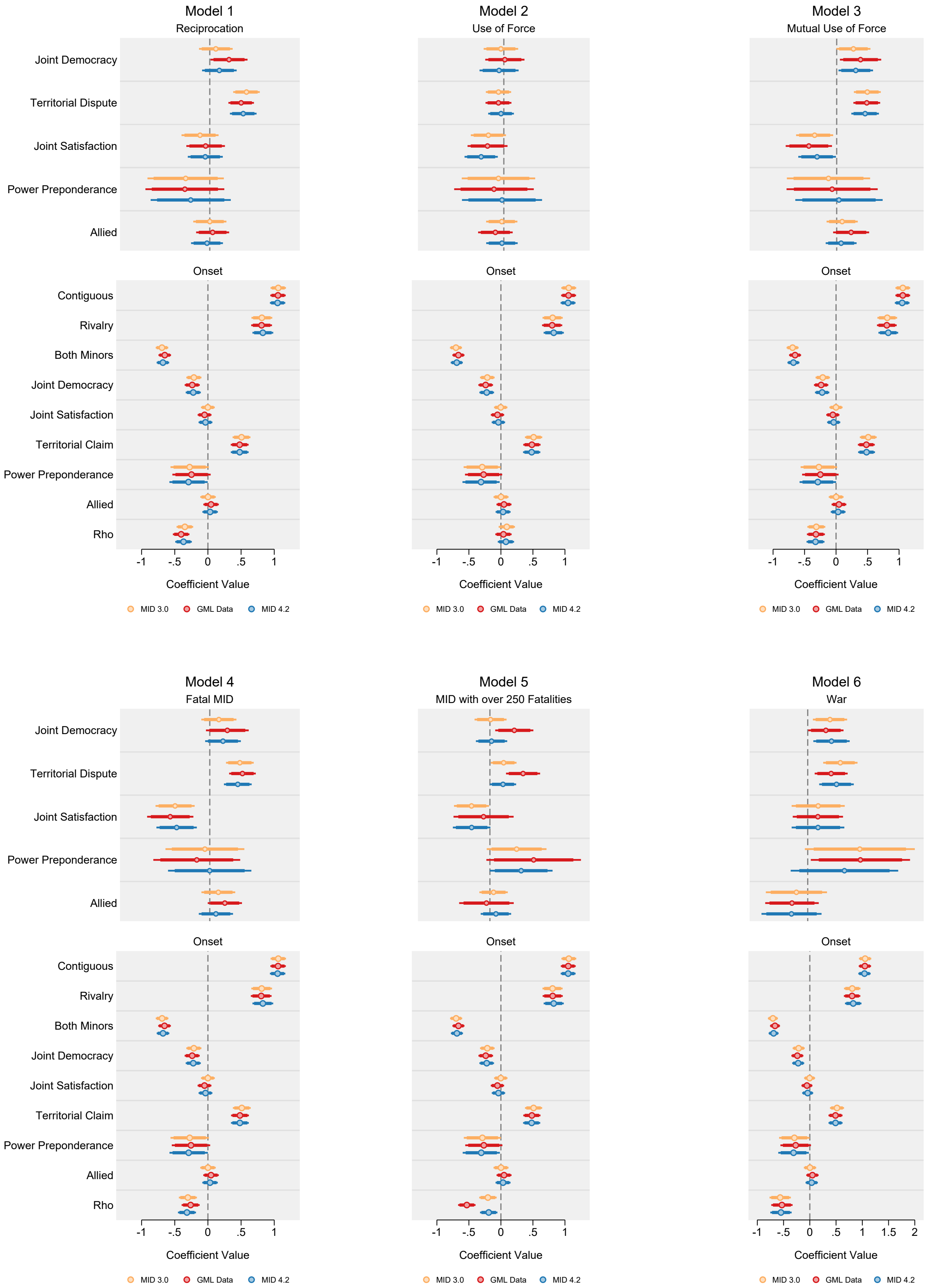

Figure 1: Replication of Braithwaite and Lemke (2011)

Note: These are selection models in which the first stage predicts onset and the second stage predicts escalation. Only the escalation equation differs between the models. The lines indicate 90 and 95% confidence intervals. Peace Year polynomials and intercepts are omitted to conserve space.

Weeks Replication

Next, we review GML's replication of Weeks (2008), whose study assesses which regime types are capable of credibly signaling resolve. Weeks examines the frequency of reciprocation in MID dyads across regime type, assuming that states sending more credible signals should experience MID reciprocation less often. She finds that democracies are no less likely to experience reciprocation than are most types of autocracies. However, she shows that personalist regimes appear to have a unique signaling disadvantage and are particularly likely to experience MID reciprocation. In GML's replication, they find that there is no significant difference between democracies and autocracies in general or between democracies and personalist regimes in particular, thus apparently overturning Weeks' key finding as well as previous findings regarding democratic credibility, i.e., Schultz (1999).

In conducting our replication, we found that GML made an error in constructing the reciprocation variable that largely accounts for the difference in results. Weeks codes MID reciprocation at the dyadic level, based on whether the target state in each individual dyad threatened, displayed, or used military force. For the years where incident-level data are available (1993-2001), Weeks takes the additional step of checking whether the target's militarized action was directed specifically toward the dyadic initiator. In contrast, GML code reciprocation at the MID level, meaning that the variable is coded as one not only if the dyadic target state reciprocated, but if *any state* that was on the target side of the MID reciprocated. Aside from not replicating Weeks' original method, this coding decision is arguably less accurate.

We re-replicate Weeks' analysis using Weeks' own method of coding reciprocation with the MID 3.02 data originally employed by Weeks, the GML data,[6] and the MID 4.3 data. The resulting analyses are reported in Figure 2. We again find few substantive differences across datasets. One of the few variables for which the various datasets do support different conclusions about statistical significance is democracy in Model 1. This variable is negative and significant using the MID 3.02 and MID 4.3 data, but falls below the significance threshold using the GML data. This suggests that the evidence of a democratic credibility advantage is somewhat fragile, as other scholars have previously found (Downes & Sechser 2012). In contrast, the positive significance of the personalist dummy is robust across all three datasets in Models 2

[6] Again, we use version 2.1 of GML's data. Earlier versions of GML's data do not include incident-level coding, making it impossible to exactly replication Weeks' method of coding reciprocation.

and 3.[7] The inconsistent results for democracy and consistent results for the personalist indicator strengthen Weeks' argument that democracies are not uniquely credible, but personalist regimes do have unique barriers to credibility. It is also notable that although GML emphasize that territorial MIDs are significantly more likely to be reciprocated in two of their regressions; this finding does not hold in any of our replication models.

The dominant conclusion from each of these replications is that the GML data do not commonly produce substantively different results in empirical analyses. We do not find that using the GML data overturns Weeks' finding about the unique credibility problems of personalist regimes or Braithwaite and Lemke's findings about the effect of joint democracy on escalation. To the extent that results do differ among the datasets, the implications are unclear, as nothing suggests that any one dataset produces findings that are more theoretically plausible than the others. However, findings that are otherwise robust to a variety of different model specifications and research designs are also most likely to be robust to variations in the MID data.

On the whole, the consistency of the results across datasets should be reassuring to MID users, as it suggests that despite some "noise" resulting from coding disagreements, we can still find the "signal," i.e., underlying patterns that are consistent and substantively meaningful. We expect that the consistency in results that we found across these two replications is also likely to hold in other studies of MID initiation, escalation, and reciprocation. Nevertheless, we caution that the results might be less consistent among studies of MID duration because of the large number of start and end date changes that GML suggest.

Conclusions and Implications

This important review of the MID data by GML highlights the fact that the MID is a complex social construct. The complicated nature of military disputes ensures that some cases will not fit neatly into one category or another. Even the most well-trained coders sometimes disagree on how the coding rules should be applied in a particular setting. This difficulty is particularly acute for historical MIDs, where current coders may not have access to the source materials used by the original coders. Nonetheless, it is reassuring that GML's recommended revisions to the MID data rarely produce substantively different conclusions in

[7] In Model 3, the significance is only at the 90 percent confidence level using the GML data.

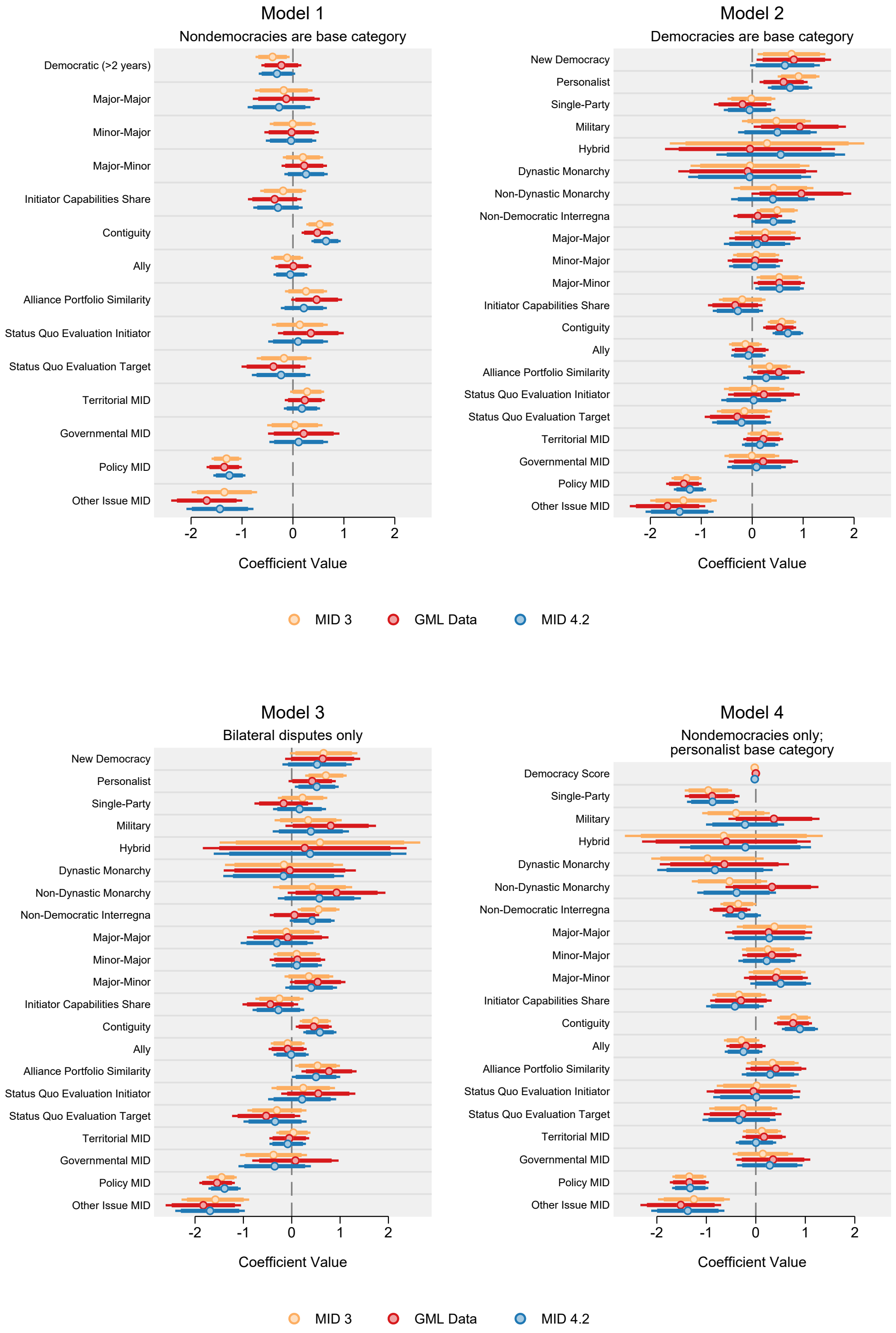

Figure 2: Replication of Weeks (2008)

Note: Logit models of MID reciprocation reproduce those in Table 4 of Weeks (2008). The lines indicate 90 and 95% confidence intervals. The coefficient for "other" regime type and model intercepts are not reported to conserve space.

our analyses.

Moving forward, the data collection practices currently used by the MID 5 project are aimed at minimizing error through the integration of text analysis and crowdsourcing techniques within the MID data collection infrastructure (Palmer et al. 2015, D'Orazio et al. 2014). First, while early iterations of the MID project relied upon non-systematized exploration of historical texts to generate the MID data, the current project gathers data through a well-defined and transparent process. Specifically, news documents are retrieved from LexisNexis using an inclusive search string. This larger set of stories is then filtered to a smaller set of stories using machine learning classifiers. These stories are read and coded into MIDs. In this way, future researchers may obtain the documents used to code MIDs, the documents ultimately judged not to contain information on MIDs, and even the documents that were retrieved but filtered out by our classifiers, determined to be irrelevant and thus not read by humans. Each of these steps enhances the ability of future researchers to reproduce and evaluate the MID data.

Second, previous iterations of the MID project relied exclusively on a small number of expertly trained researchers to code the news stories. By integrating crowdsourcing techniques and increasing the overall efficiency of the data collection process, we increase the number of individuals who read any given document. Utilizing multiple coders increases the ability to identify news stories containing information about militarized events, thus minimizing the possibility of errors that result from coder fatigue and idiosyncratic decisions. This approach enhances validity and reproducibility by ensuring a larger number of coders contribute to coding decisions (Benoit et al. 2016). Nevertheless, in the MID 5 process, expert coders still make all final classification decisions.

In short, the data collection strategies of the MID project have evolved with our ability to collect, store, and classify information on international conflict events. We are hopeful that these strategies will reduce the potential for future difficulties when attempting to recover or reproduce information contained within the MID data.

Supplemental Information

Replication files and an online appendix explaining our response to each of the recommended MID drops are available at http://personal.psu.edu/rum842/ and at the *International Studies Quarterly* data archive.

References

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver & Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–295.

Braithwaite, Alex & Douglas Lemke. 2011. "Unpacking Escalation." *Conflict Management and Peace Science* 28(2):111–123.

D'Orazio, Vito, Michael R. Kenwick, Matthew Lane, Glenn Palmer & David Reitter. 2014. "Crowdsourcing the Measurement of Interstate Conflict." *PLOS ONE* 11(6):e0156527.

Downes, Alexander B. & Todd S. Sechser. 2012. "The Illusion of Democratic Credibility." *International Organization* 66(3):457–89.

Gibler, Douglas M. 2018. *Militarized Interstate Dispute Narratives, 1816-2010*. Lanham, MD: Rowman and Littlefield Publishers.

Gibler, Douglas M., Steven V. Miller & Erin K. Little. 2016. "An Analysis of the Militarized Interstate Dispute (MID) Dataset, 1816-2001." *International Studies Quarterly* 60(4):719–730.

Hui-wen, Jen. 1993. "China; Paper Says, 'Combat Readiness' in China is Response to Political Change in Taiwan." *BBC Summary of World Broadcasts* February 5.

Palmer, Glenn, Vito D'Orazio, Michael R. Kenwick & Matthew Lane. 2015. "The MID4 Data Set, 2002-2010: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science* 32(2):222–242.

Schultz, Kenneth A. 1999. "Do Democratic Institutions Constrain or Inform? Contrasting Two Institutional Perspectives on Democracy and War." *International Organization* 53(2):233–266.

New York Times. 1955. "No Peru Army Found on Ecuador Border." September 15.

Weeks, Jessica. 2008. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization* 62(1):35–64.