Responses to Reviewers' Comments for Manuscript 2023-07372

Estimating Conflict Losses and Reporting Biases

Addressed Comments for Publication to

PNAS

by

Benjamin J. Radford, Yaoyao Dai, Niklas Stoehr, Aaron Schein, Mya Fernandez, and Hanif Sajid

To the editor,

Thank you for giving us the opportunity to respond to Reviewer 2's remaining concerns. We will do our best to address these concerns in this letter. We would also like to thank Reviewer 2 for their positive assessment of our work. While Reviewer 2 has not suggested any manuscript changes, we have tried to amend our manuscript to clarify the concerns raised by Reviewer 2. We acknowledge Reviewer 2's skepticism of Bayesian methods, either in this particular setting or in general. While we think a broad segment of the PNAS readership will not share this skepticism, and we do not feel we have the space in the brief report to address foundational concerns, we do want to make sure that we alleviate this concern to the best of our ability in this letter.

The crux of Reviewer 2's concern seems to come down to whether or not a Bayesian model is an appropriate alternative to an OLS or MLE-based ("Frequentist") fixed effects (FE) estimator. However, as the reviewer points out, these are not mutually exclusive. One can estimate a FE model via Bayesian inference by placing an improper prior on the fixed parameters. *We have now estimated such a fixed effects model precisely as the reviewer suggested here and in their previous review.* We are satisfied that the results are substantively comparable to those that we report in our manuscript though, for the reasons noted in this letter, we think one should prefer the approach presented in the manuscript.

While we are unsure whether the reviewer's primary concern is with Bayesian inference itself or with the choice to use RE over FE, we believe we have addressed *both* possible concerns. With respect to the latter, we have re-estimated our model with FE. With respect to the former, we have elaborated upon our need to use a Bayesian model and explain, in detail, why the Bayesian RE model should be strongly preferred to a Frequentist or FE alternative in this particular case. We summarize both efforts for your convenience.

Fixed Effects Robustness Test. A core component of Bayesian modeling is the requirement that priors be defined for all parameters. These priors are typically probability

2

distributions. Using probability distributions as priors guarantees an identifiable posterior model. It is considered good practice to allow prior knowledge or assumptions about the data generating process (DGP) to inform the choice of priors. One can also place improper priors on parameters. For instance, a uniform prior that places equal probability on *all* real values is an improper prior since it does not integrate to 1. Placing such a prior on our bias parameters is equivalent to bias fixed effects. As was suggested by the reviewer, we have placed an improper (FE) prior on $\gamma_{st}^{\text{bias}}$ and re-estimated our model. We are very happy to report that the results are comparable to those presented in our manuscript. We go into more detail in our response to the reviewer.

Why a Bayesian Model? Nonetheless, we still believe that our original Bayesian RE model is superior to a Bayesian FE model and that no purely Frequentist alternatives exist. We are therefore confident that we have selected the best model for our manuscript from among all possible alternatives of which we are aware. In our response, we do the following with respect to justifying our model choice:

1. Explain why a simple OLS/MLE FE model is not suitable for our problem (though Reviewer 2's intuition about the similarity between fixed intercepts and our bias terms is correct).

2. Explain why, even if an OLS/MLE FE were an option, it would not be identifiable (due to the perfect multicollinearity between reporting sources).

3. Explain why the typical remedy to the above problem, omission of a single reporting source bias term (which we do in our Bayesian FE model), would introduce into our analysis an insidious hidden assumption that is not obvious, arguably much stronger than the assumptions imposed by our priors, and would change the interpretation of our model.

4. Explain that our choice to use Bayesian inference rather than Frequentist methods is motivated by other modeling concerns, not by our desire to place priors on the bias parameters. Rather, we chose to use RE as a consequence of choosing to use a Bayesian model. We needed a Bayesian model so that we could properly handle the missingness in our dataset and the relatively complicated latent variable estimation.

5. Explain that we did not choose our model because we believe it to be better than existing alternatives, but because *there are no existing alternatives* of which we are aware. It is entirely possible that better models of latent conflict losses, including models based on FE, will be discovered in the future and we encourage researchers to work toward this end – we think this is an important problem! However, we are not aware of any existing models that encode our assumptions about the DGP and therefore we made one. To summarize, our model was made to accommodate the following issues and assumptions:

   - Observations at multiple time units (daily / cumulative)

     - **Solution**: our model is multivariate in the sense that it has two dependent variables (daily and cumulative observations), each of which has its own functional form. However, parameters are shared across these two functional forms.

   - Observations of point values as well as range values ("between X an Y")

     - **Solution**: We control for reports being low estimates, high estimates, or neither (i.e., point estimates).

   - Observations from correlated time series

     - **Solution**: We draw coefficients for our latent variables from a multivariate normal distribution that allows for time series to be correlated with one another (e.g., armored vehicle losses are positively correlated with troop losses).

   - Observations from multiple reporting sources with a hierarchical correlation structure and high data sparsity

     - **Solution**: We control for source-target reporting variations and source-target-category reporting variations using multilevel effects. These are the "bias" terms and they are analogous to fixed or random intercepts.

   - Unobserved ("latent") daily loss values which are proportional to observed reports, conditioned on the above items

- **Solution**: We model latent daily losses using target-category intercepts, trends (slopes) and cubic B-splines. These are then multiplied by the scalar bias parameters and other relevant model parameters.

- Multiplicative effects rather than additive effects. We assume that biases scale values; they don't shift them by a constant.

  - **Solution**: Our model is log-linear in parameters. We estimate the parameters in log space and add them to the logged latent loss values. These values are then exponentiated before we compute the model likelihood. Therefore, the effects of our parameters are multiplicative, not additive.

- None of the observed time series are complete (high levels of missingness).

  - **Solution**: We estimate a complete set of latent time series in the presence of missingness by allowing our daily latent variables to be correlated across loss categories, thereby sharing information from the time series for which observations are available. We also use the observed cumulative values to inform our estimates of daily losses, even for cross sections within which there are no reports of daily losses.

We are very hopeful that the new FE model, manuscript revisions, and responses found in this letter will alleviate the reviewer's remaining concerns. This revision process has been very helpful for us and the current manuscript is more thorough and more clear as a result. We are very appreciate of both reviewers' comments.

Please find enclosed the revised version of our previous submission entitled "Estimating Conflict Losses and Reporting Biases" with manuscript number 2023-07372.

Sincerely,

The authors

Authors' Response to Reviewer 2

We want to thank the reviewer once again for their careful reading of our manuscript, their supportive comments, and their deep engagement with our research methodology.

We first wish to summarize our response to the reviewer's primary concerns. We chose to develop a Bayesian model over using an existing Frequentist/MLE model because we do not believe that any other Frequentist (or even Bayesian) models exist that are suitable for our problem. This is distinct from our choice to use RE over FE for the bias parameters. We chose RE because they provide us with a clearer interpretation of our results and they are a natural choice when using a Bayesian model. In practice, however, we show here that the RE and FE parameterizations of our model produce very similar results.

We also want to note that the overall goal of our model is estimate conflict losses under certain, reasonable assumptions, and that doing so requires us to accomplish many things simultaneously: estimating correlated latent variables over time, in the presence of missing data, and while accounting for systematic variations (biases) across data sources. This new model is only one of several contributions we make in this brief report, which also include the production of a dataset on conflict loss reports in Ukraine and the presentation of a new research question: how to measure conflict losses from contemporary loss reports that are noisy, biased, and often contradictory of one another.

While we will address all of the reviewer's comments in order below, we first want to highlight a new robustness test that we believe satisfies the reviewer's major concerns. Specifically, the reviewer suggests "group-specific [fixed] intercepts...using either MLE or Bayesian methods." We have successfully estimated our model with fixed bias terms rather than random bias terms. We do this by imposing, at the reviewer's suggestion, an improper prior (Uniform$(-\infty, \infty)$) on $\gamma_{st}^{\text{bias}}$. Because the quantities of interest are actually the second-level bias terms, $\beta_{st,c}^{\text{bias}}$, we allow these to be distributed Normal$(\gamma_{st}^{\text{bias}}, \sigma)$. This new model perfectly satisfies the reviewer's suggestion (from the first round of revisions): "an alternative model specification in which the mean bias terms (γ_{st}^{bias}) are fixed effects." This model is given below, with the improper Uniform prior shown in Eq 12.

$$y_i^{\text{daily}} \sim \text{Pois}(\exp(\mu_i^{\text{daily}})) \tag{1}$$

$$y_j^{\text{cum}} \sim \text{NB}(\exp(\mu_j^{\text{cum}}), 1/\exp(\phi_{ct[j]})) \tag{2}$$

Regression Means

$$\mu_i^{\text{daily}} = \theta_{ct[i],d[i]} + \beta_{c[i],st[i]}^{\text{bias}} + \beta_{s[i]}^{\min} I_i^{\min} + \beta_{s[i]}^{\max} I_i^{\max} \tag{3}$$

$$\mu_j^{\text{cum}} = \ln(\Sigma_{k=1}^{d[j]} \exp(\theta_{ct[j],d[k]})) + \beta_{c[j],st[j]}^{\text{bias}} + \beta_{s[j]}^{\min} I_j^{\min} + \beta_{s[j]}^{\max} I_j^{\max} \tag{4}$$

Latent Time Series

$$\theta_{ct,d} = (B\beta_{ct}^{\text{spline}})_d + \beta_{ct}^{\text{const}} + \beta_{ct}^{\text{trend}} (d/365) \tag{5}$$

Priors

$$\beta_c^{\text{const}} \sim N(\mu^{\text{const}}, \sigma^{\text{const}}) \tag{6}$$

$$\beta_{ct}^{\text{trend}} \sim N(\mu^{\text{trend}}, \sigma^{\text{trend}}) \tag{7}$$

$$\beta_{ct}^{\text{spline}} \sim N(0, \Sigma^{\text{spline}}) \tag{8}$$

$$\beta_s^{\min} \sim N(\mu^{\min}, \sigma^{\min}) \tag{9}$$

$$\beta_s^{\max} \sim N(\mu^{\max}, \sigma^{\max}) \tag{10}$$

$$\beta_{c,st}^{\text{bias}} \sim N(\gamma_{st}^{\text{bias}}, \sigma_{st}^{\text{bias}}) \tag{11}$$

$$\gamma_{st}^{\text{bias}} \sim \text{Uniform}(-\infty, \infty) \tag{12}$$

$$\phi_{ct} \sim N(\mu^{\phi}, \sigma^{\phi}) \tag{13}$$

As should be expected, many bias terms obtain greater CIs in the FE model than in the RE model; this is of course the result of some source-target-category cross-sections containing very few data points and was the motivation for using multilevel (partial pooling) bias terms in the first place. More importantly, as we explain below, the bias terms must now all be interpreted *relative* to the omitted source bias category (as is common in FE estimators) and the loss values are no longer interpretable as expected unbiased losses. However, the FE model is substantitively similar to the RE model. We

discuss these results in response to Comment 3. Nonetheless, for the reasons we detail throughout this letter, one should still prefer the proposed model to the FE alternative.

Comment 1

The authors have more than sufficiently addressed my minor concerns.

We are very glad to hear that the reviewer is satisfied with most of our revisions and clarifications. It is very important to us that we thoroughly address all concerns, no matter how minor, and we are grateful to the reviewers for bringing them to our attention. As you might imagine, drafting the previous response letter was a learning experience for us as well and we were relieved to find that our robustness tests in response to the reviewers' comments left us even more confident in our original modeling choices.

Comment 2

As for my major concern, I was particularly worried about how the statistical model accommodates "source-specific biases" in the reporting of battle losses. I was worried about the sensitivity of the paper's conclusions - in particular those about the degree to which Russia and Ukraine over- or under-estimate battle losses - to distributional assumptions imposed on biases across sources.

In response, the authors pursued two approaches. First, they clarified that their statistical model only imposed a prior (normal with mean zero and an estimated variance) on the distribution of source-specific biases. Along these lines, they changed some terminology used to explicate the statistical model, changing "random effects" to "coefficients" or "multi-level effects". Second, the authors change how they parameterize this prior and show that the estimated empirical distributions of the source-target bias terms are not wildly sensitive to these changes.

The authors have addressed my major concern to some degree: they show that some results remain the same under different distributional assumptions. As such, my general assessment of the work remains positive.

We thank the reviewer for their kind assessment of our manuscript and hope that we can adequately alleviate their remaining concerns.

Comment 3

However, their response raised two issues:

1. How do the substantive effects (reported in lines 125–138 of the updated manuscripts) change across these models? E.g., is the rate of Russian misreporting ("for every loss suffered, Russian sources report only 0.3 losses) the same across these models? Same for the reported rate of how the Ukrainian sources overestimate Russian deaths?
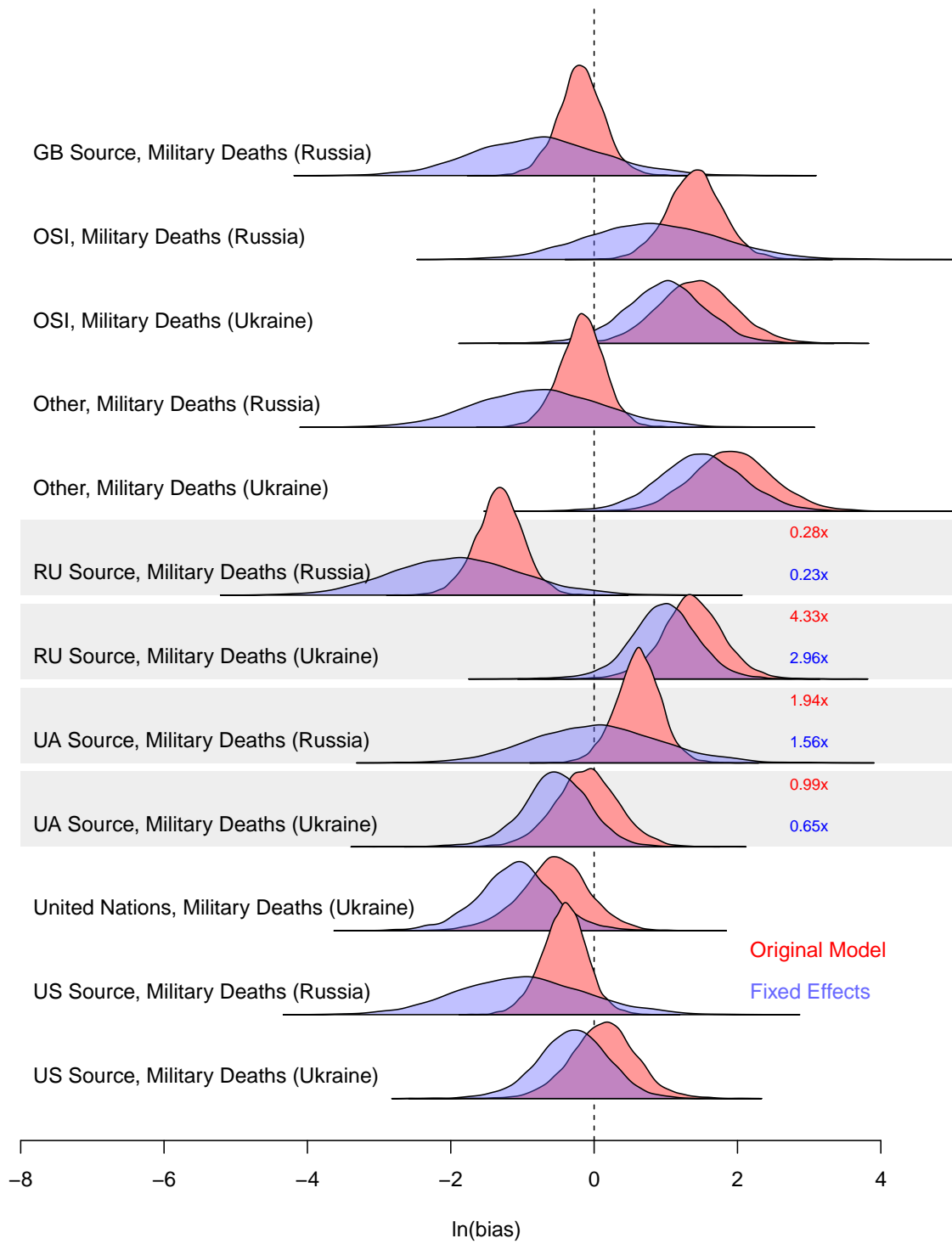
Figure 1: Fixed effects biases, log scale. Values referenced in the manuscript are high-lighted. exp(bias)-transformed values are given to the right of highlighted distributions.

We have run *both FE and RE versions of our model*, as requested, and the results are substantively similar. All parameters mentioned in this comment (and more) are compared across FE and RE models in Figure 1. As we should expect, the RE specification is more efficient due to partial pooling. Also, all parameter values are shifted by a constant in the FE model relative to the RE model. This is not surprising because the interpretation of FE is always in comparison to the omitted ("baseline") category (Ukrainian reports of Ukrainian losses).[1] Nonetheless, the bias estimates are substantively similar in both absolute and relative terms.

Specifically, we find the RE interpretation of Russian misreporting ("for every loss suffered, Russian sources report only 0.3 [rounded] losses") would become, under FE assumptions, "for every loss suffered, Russian sources report only 0.2 [rounded] losses."[2] Our estimate of Ukrainian overestimation of Russian deaths decreases from 1.94 to 1.56. As can be seen in the figure, these values are well within each other's confidence intervals. In the figure we highlight all bias parameters referenced in the manuscript.

Comment 4

2. The paper claims to make a methodological contrition by "provid[ing] a novel approach to measuring casualties and fatalities given multiple reporting sources and, at the same time, accounting for the biases of those sources" and by "dealing with source-specific biases in a principled way, treating them as parameters to be estimated."

We thank the reviewer for pointing this out and, though we don't think it is the intent of the reviewer to suggest rephrasing this portion of the abstract, we have decided to do so:

---

[1] We elaborate further in response to Comment 5.

[2] Again, the second value here cannot be interpreted as an absolute bias but as a bias relative to the "known" zero bias of Ukraine, encoded by our omitted category. Also, the rounding does cause the difference to appear greater than the model actually reports.

"In this paper, we offer an approach for measuring casualties and fatalities given multiple reporting sources and, at the same time, accounting for the biases of those sources."

Comment 5

Yet, I am having a hard time understanding the strengths and weaknesses with this new approach, especially regarding how scholars might use it to estimate source-specific biases. In their response, the authors write that "a traditional fixed effects parameterization of this particular model (such as an MLE FE estimator) would likely not be identified." To what degree is this a weakness? Or something we should be concerned about?

...

I think the analogy to fixed effects estimation was useful but highlights the issue: I can fit a linear model to panel data including group-specific intercepts ($y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$) using either MLE or Bayesian methods. I don't need Bayesian methods to identify or estimate the group specific intercept, α_i.

To what degree is this a weakness? Or something we should be concerned about? We do not see this as a weakness. We have now implemented a (Bayesian) FE version of this model and shown that it produces similar results to the RE version. The FE alternative is, indeed, identifiable under the common FE strategy of omitting a "baseline" or "reference" category.[3] However, we remain confident that the approach described in the manuscript is superior to the FE alternative for multiple reasons that we outline below.

On the identifiability of FE and RE models: both can be identified under different

---

[3]In circumstances where groups are mutually exclusive and exhaustive, FE can be estimated through the omission of one group. Failure to do so results in perfect multicollinearity between groups.

conditions. Our RE model is identified under modestly informative and conservative priors while the FE model is identified by the omission of a "baseline" category. In Bayesian inference, assuming proper posterior distributions, there is no such thing as a non-identifiable model. Andrew Gelman writes: "In the broadest sense, a Bayesian model is identified if the posterior distribution is proper. Then one can do Bayesian inference and that's that. No need to require a finite variance or even a finite mean, all that's needed is a finite integral of the probability distribution."[4]

I can fit a linear model to panel data including group-specific intercepts...using [MLE]... We are confident that this is incorrect in this particular case.[5] We will explain our concern with this misunderstanding by highlighting three points:

1. You could fit an MLE FE model and interpret the fixed intercepts as source biases if you had: a complete panel, no latent variables, an a priori preferred (known unbiased) baseline category, and no need to compute monotonic cumulative sums of the latent variables. Our choice to utilize a Bayesian model with multilevel effects (or, in the first response letter, single level effects) is motivated primarily by our need to accommodate (1) in situ imputation of missing data, (2) cumulative summation of latent random variables, and (3) multiple link functions for two different outcomes (daily and cumulative losses). The selection of RE did not prompt us to use Bayesian methods; our need to use a Bayesian model motivated (in part) our use of RE.[6]

2. However, the assertion that the model could be estimated via FE is not as helpful as it first sounds. The FE estimator implicitly encodes an assumption about the

---

[4] https://statmodeling.stat.columbia.edu/2014/02/12/think-identifiability-bayesian-inference/

[5] The reviewer is in good company here because we also spent a fair amount of time pondering these identifiability problems early in this project.

[6] The closest Frequentist approach we can imagine for this problem is to first impute the missing data and then to use something like an item response theory model to estimate the biases and latent loss values. However, this is already more complicated than specifying our own model and we would still need a custom model to account for the different functional forms of the daily and cumulative dependent variables. Furthermore, most multiple imputation methods are themselves Bayesian or pseudo-Bayesian, requiring draws from a posterior predictive distribution, and therefore even this "Frequentist" approach is not as Frequentist as it seems.

biases that we believe is inappropriate and, in fact, much stronger than the assumptions we encode in our original priors. Specifically, all reporting sources are mutually exclusive and therefore one source bias term would necessarily need to be omitted from the model (a "baseline" category in FE terms – otherwise the model would not be identifiable!). Doing so in this case is equivalent to asserting a zero bias with no uncertainty for that particular source. We think this is misleading: omitting one reporting source category would have a dramatic impact on the interpretation of the model yet would appear perfectly normal to those who are only familiar with FE in the context of OLS or MLE regression estimators.

We very much prefer placing a modestly informative (and conservative) prior of zero mean on the distribution of source biases because doing so does not require us to (a) select a preferred unbiased source to act as the baseline omitted category or (b) assume that any single bias term is precisely zero. This is especially important for us because our objective is to obtain expected values for losses conditioned on the estimated biases. If we use a FE estimator, we are not only asserting a zero bias for the omitted reporting source but our estimated losses would then necessarily be the expected losses for that reporting source (because we have by omission encoded with perfect certainty that the source's bias term is known to be zero). They could no longer be interpreted as the expected losses conditioned on all estimated source biases (i.e., "unbiased" estimates).

The zero mean prior does not enforce any particular value on the biases but it *does* encode an assumption that there are no biases that affect *all* reporting sources. On the one hand, this may sound like a strong assumption, but it is analogous to the common assumption that our data do not suffer from systematic measurement error. This assumption is so common that it goes unaddressed in nearly all scientific papers. However, the cool thing is: if researchers were to discover that all reports suffer from a systematic bias, they could simply encode that knowledge in this prior by adjusting the prior mean accordingly! In fact, we did precisely this in our first response letter when we put a hyperprior on the mean value of γ^{bias}.

However, as we explained in our previous response, we have no reason to expect

this systematic measurement error. Furthermore, this assumption is common to most (perhaps all?) measurement models (e.g., item response theory models) which, when it really comes down to it, are often just fancy weighted averages.

3. We are concerned that this response suggests a preference for Frequentist methods that is unjustified. That a researcher *can* estimate a model via MLE (which, we reiterate, we believe cannot be done in this case) does not preclude the possibility that this or a similar model could *also* be estimated via other means. Bayesian inference gives us the flexibility to carefully represent our theorized DGP as a generative model without deriving novel MLE estimators or relying on pre-packaged regression models, none of which are, to the best of our knowledge, suitable for our data. We are not suggesting Bayesian inference is inherently better – it is just more suitable for our particular task and data.

> ### Comment 6
>
> Can any prior on γ_{st}^{bias} or $\beta_{c,st}^{bias}$ be used even one that is very diffuse or one that is improper?

Can any prior...be used? Diffuse priors are fine. In fact, we would argue that our priors are diffuse. In both our manuscript model and our robustness test models from the first response letter, we place a hyperprior on the variance of the γ_{st}^{bias} bias terms. This hyperprior is drawn from a standard normal distribution and exponentiated. If, for example, we obtain a draw from the normal distribution of 2 $(0 + 2\sigma)$ this corresponds to a prior standard deviation for γ_{st}^{bias} of $\exp(2) = 7.4$. If we were to then draw a $\beta_{c,st}^{bias}$ two standard deviations above the mean (0) of γ^{bias}, this would be ≈ 15. Because the parameters are log-linear, this would correspond to a $\exp(15) \approx 3,000,000\times$ bias, far above any reasonable expectation for bias values.

...or one that is improper? We successfully use an improper prior in our FE model described above. This is a maximally diffuse prior on the biases. However, this

change comes with caveats. To avoid non-identifiability, we must omit one bias term (as we discuss above), but doing so essentially invalidates the presumed benefits of using an improper prior: we've just shifted our explicit assumption about a prior mean into an implicit assumption about a bias term's actual "known" value. Secondly, improper priors are difficult to justify in a Bayesian setting. If, for example, we believe that a $3,000,000\times$ bias term is absurd in reality, then we must concede that an improper prior is unjustified: it would place equal prior weight on this (and more!) extreme values as it does on nearer-zero values. Doing so would be throwing away valuable data that we possess about the reasonable scale of reporting biases.

Nonetheless, we want to emphasize that we did successfully estimate our model with an improper FE prior on γ^{bias} and produced substantively similar results to our original model.

Comment 7

Part of this is that I do not use/implement multilevel models in my own research. Another part is that the space constraints are a bit limiting. But a substantial part is that the paper is suggesting that its approach can be used to estimate source-specific biases without a "true" account of the battle losses. This seems like an enormous contribution with many applications, but I still do not understand the assumptions or data required to identify and estimate these biases.

Thank you for the very kind assessment of our contribution. The primary assumption that we need to make is to "center" our bias terms. There are at least two ways to accomplish this: one is to impose a soft-centering prior[7] and the other is to select one bias term to represent a "baseline" against which the others are evaluated. In our model, the soft-centering prior takes the form of a mean zero prior on the source-target bias terms (γ). This gives us a convenient interpretation: we are making the assumption

---

[7] https://mc-stan.org/docs/stan-users-guide/parameterizing-centered-vectors.html

that the source-target bias terms (whatever they may be) are centered around zero in expectation or, in other words, that there isn't a systematic bias that affects all of these terms on average. An equivalent way of looking at this prior is that we are assuming there is not systematic measurement error in the loss reports.

We have attempted to make this more clear in the manuscript as follows. Of course, the strict three-page limit means that we cannot be quite as detailed in the manuscript as we are in this response letter.

> Every observation is scaled by a multilevel bias coefficient that is specific to its source-target pair (indexed by st) and category, $\beta_{c,st}^{\mathrm{bias}}$, to account for systematic over- or underestimation. These are normally distributed around source-target specific means which are themselves normally distributed with prior mean zero (Eq 9). We assume zero-centered bias terms, encoding the conservative belief that a source is unbiased absent data to indicate otherwise. A non-zero mean would encode belief in a systematic bias (e.g., systematic measurement error). We model bias terms hierarchically to mitigate class imbalances through partial pooling. When estimating losses, we set the bias terms to zero (i.e., "no bias").

In choosing a multilevel structure for the bias terms, we are assuming that source-target-category bias terms are correlated within source-target groups. In other words, we are assuming that the `Ukraine-Russia-Tanks` bias and the `Ukraine-Russia-Aircraft` bias are likely to be drawn from the same distribution centered around a mean `Ukraine-Russia` bias. This is a common assumption in multilevel models and is especially useful in cases of data imbalance or data sparsity.

As a methodological contribution, we describe for researchers techniques that we found helpful in modeling conflict losses. These techniques are described in the Materials and Methods section of our manuscript and will be available in the replication archive that we will publish. The replication archive contains the R code and Stan code to reproduce the model from the manuscript as well as all 13 robustness tests. We believe that the combination of these techniques is unique and that their application to this

problem is novel. We do not want to oversell this contribution: it is not that we believe our model is better than competitor models but instead that *there are no* previous models that are suitable for the data that we have collected.

We very much hope that the reviewer is right and that other researchers will find more applications for this (or similar) models. While our focus here is on measuring conflict losses in the presence of reporting biases and incomplete panel data, we can imagine that similar techniques could be applied to any number of processes in which measurement is difficult and data are noisy/missing.