Author response to reviews of

Manuscript Manuscript number

Moral Foundations of U.S. Political News Organizations

Padfield, Buchanan, & Jordan
submitted to *Meta-Psychology*

[RC] Reviewer comment Manuscript text

Dear Dr. Elson,

Thank you very much for taking the time to consider our manuscript for publication at *Meta-Psychology*. Because we made a large number of changes to the document as requested, we do not copy all the changes here, but note a few specific areas. We believe the manuscript is much improved and thank the reviewers for their thoughtful comments.

In the following we address your and each reviewers' concerns point-by-point.

Preprint: `https://osf.io/rmj73/`

OSF: `https://osf.io/5kpj7/`, `https://osf.io/rmj73/`

1. Reviewer #1

[RC 1.1.] **This manuscript reports results from a pair of studies examining the validity of the Moral Foundations Dictionary (MFD), specifically by comparing use of terms from the MFD by previously-identified conservative and liberal news sources. The first study compares MFD term usage in two liberal (NPR, NYT) and conservative (Fox News, Breitbart) news sources, finding no significant differences in MFD words used for any of the five moral foundations dimensions between the two types of news sources. A second study, its design based on part on lessons learned from issues with data collection from the first study, did the same with twenty news sources and specifically with coverage of Brett Kavanaugh's Supreme Court of the United States nomination and confirmation and the 2018-2019 U.S. government shutdown stemming from disputes about a proposed U.S.-Mexico border wall. The second study found a significant (albeit small, effect-wise) difference between liberal and conservative sources only for the harm-care dimension, as the manuscript notes has been found in previous research, but only for the Kavanaugh coverage. Results are interpreted as casting further doubt on the utility of the MFD, and more conceptually the applicability of the MFT dimensions, to explain partisan political differences. I appreciate the level of detail in the description of the studies' method and analysis, including unexpected issues such as the challenges in scraping some of the first study's sources. The thorough and highly organized presentation of open materials and the clear explanation of methods including use of specific R packages provides useful guidance to both critical examination of these studies and future research building on them.**

The review of research behind the MFT, MFQ, and MFD is parsimonious but thorough, and in my

opinion quite reasonable about the shaky conceptual foundations of all three as well as the limited empirical evidence for reliability and validity of the measures. If anything, the review of the literature is forgiving given how often measures like MFT do not even demonstrate adequate inter-item reliability. In any case, the rationale for the study is clearly justified by a fair review of a popular conceptual framework that very much has feet of clay. I also find the justification for the selection of sources to be adequate, and appreciate the authors' basing them in the Mitchell et al. article rather than simply relying on an intuitive assumed rationale for the selection.

Thank you for your kind words!

[RC 1.2.] If there is more systematic justification for why these precise four news sources were chosen over all other options for Study 1, it might be good to share. If there is no such specific justification, though, I don't think it is a problem that needs addressing further as there is adequate rationale for these two pairs of sources to be—at the very least—fine examples of the two source variable conditions. (Although a critique might be that Breitbart is arguably less mainstream than the other three sources, though that doesn't seem to invalidate the study's findings.)

As described in text, mostly these were chosen due to perceived and self-described political lean. However, we added a sentence about how they also practically worked best:

"At the time of our data collection, these sources were also selected because they were open (i.e., no subscription required to access articles), and their websites were designed in a way that made webscraping possible."

[RC 1.3.] Given that in many ways the first study ends up being something of a de facto pilot study for the second (which I believe adds substantially to the credibility of the latter as well as the message of the pair as a package here), this is not a major issue as five times as many sources are used in the second study with similar findings. The manuscript seems to imply that these twenty sources represent the top ten sources from each valence of lean in the Mitchel et al., article; if that is so, it may be good to state as much more clearly. If I misunderstood and more arbitrary criteria were used to select each set of ten sources from the Mitchell et al. article's results, that may need to be clarified as well.

We picked the ones based on Mitchell's article that also were practically usable. Therefore, we added this sentence to our second study source section as well:

"These sources were primarily selected due to their political lean but also due to our ability to webscrape their contents without a subscription based service."

[RC 1.4.] Again, though, I do not perceive these as concerns even if there is no more specific justification given that these seem appropriate conditions for studies viewed as experiments rather than as some sort of representative source sample. That said, I am reluctant to consider these studies simply experiments per se rather than perhaps as natural experiments. I am not sure I am correct, as I have limited experience with the latter, but I wonder if it might at least be worth suggesting renaming both studies as natural experiments rather than solely as experiments. That said, I defer to the editorial staff, reviewers, and authors themselves regarding whether the manuscript would be more accurate continuing to describe these studies as experiments or if it might be a good idea to call them natural experiments.

We understand the point of view of this critique and have changed the language accordingly throughout the manuscript.

[RC 1.5.] **As a minor point, given that the studies are described as "a further conceptual replication of Frimer (2020)," which seems apt (as well as some of the measures and interpretation thereof are also based on Frimer), it may be useful to describe the findings thereof for the reader at least briefly (even if only in conceptual terms). The study is mentioned parenthetically in a sentence describing previous literature, but a sentence or so more about the specific contribution of Frimer (2020) might aid the reader given it is described as uniquely central to these studies' motivation and design.**

The section in the literature review describing the Frimer study has been expanded to fully describe the study.

[RC 1.6.] **While the manuscript is well-written, there is at least one error I would consider very important, namely at least one misspelling of Brett Kavanaugh's surname. It would, of course, be a good idea to take a very close look for such errors, even if they are not technically germane to the credibility of the studies' design and results. I believe that addressing these issues will result in a useful contribution to the literature about a theoretical framework and measure that seems increasingly in need of some reconsideration—particularly given its popularity.**

The misspelling has been fixed.

2. Reviewer #2

[RC 2.1.] **The present manuscript applies a version of the moral foundations dictionary to news articles. I have to disclose that I am quite strongly predisposed against the validity and generalizability of MFT overall, first because it is extremely US-centric, second because I have seen it fail quite spectacularly in survey applications. I am also quite skeptical of dictionary-based approaches overall. The negative findings of the paper were comfortably confirming my expectations, which might positively bias my review.**

I have several major and minor concerns:

(1) **I don't exactly understand the purpose of the paper: Is it a validation study of the MFD or an attempted application of MFT to the analysis of media content? This concerns the interpretation of the findings: Do we believe that moral foundations simply did not impact the coverage of the issues, or just that the MFD did not pick up the latent messages because it is a flawed instrument? Would the cited improved methods such as manual coding or modern large language models be more successful? We do not know because the MFD is not systematically compared to alternative approaches. Ironically, the authors mention the multi-trait-multi-method approaches related to MFT, but do not apply them themselves. However, applying on method to one sample, tells us little about the validity of the method *or* the idiosyncrasies of the sample. Therefore, I am unsure what we can learn from the paper.**

The cited manuscript from our original paper was another manuscript the team was working on. We decided to merge those results with this manuscript for stronger evidence about the issues with the MFD. We have rewritten sections to clarify, using the theoretical information proposed by the MFT and other publications, we examined if those results could be found: within several different types of sources and with potential improvements to the MFD. Study 1 now includes our MTMM results.

[RC 2.2.] (2) **The authors extend previous research on the MFD by "augmenting" the word counts with scores from another dictionary (Warriner et al.) I don't have access to the cited papers, especially Frimer's, but if this is an original idea by the authors, it needs *much* more theoretical and methodological justification. Previous applications of the MDF have by design only tapped the salience of the different dimensions in different texts, and it's not obvious to me whether extending this to valence is as straightforward as the present paper suggests. And even if one accepts the premise that context-free ratings of word lemmas by Mturk Workers are useful measures of some kind of word "valence", it's the authors burden of proof that multiplying one set of scores with a completely different set of word scores from another task makes sense. At the very least, I'd expect a quasi-replication study to start out with the traditional application of the MFD, i.e. simple word counts/relative frequencies as outcomes, and then check if and how multiplying them by z-standardized (why btw?) valence scores changes things.**

In our revisions, we have now tested several ideas: 1) if we add more words can we see how the MFD is related to the MFQ (which has good validity evidence)? 2) if that doesn't work, can we augment scores with valence to improve the results? The original MFD was created by simply thinking about words that should relate to the moral foundation areas, and we expand by suggesting new terms through free association and written texts focusing on moral areas. While some frequencies correlate with MFQ results, the relation to the MFQ does not appear to hold with MTMM results.

We examined valence to improve MFD scoring by adding context to the words included. Within one foundation, concepts are both positive and negative. We examined if adding that information would improve the ability to find expected differences. We added citations for justification for valence ratings, as this is very popular technique for sentiment analysis. As noted in text, we z-scored to help with interpretation, as the dataset ranges from 1 to 9, but -1 to 1 is easier to interpret (i.e., greater than zero is positive, less than zero is negative). For comparison, we also provide the results of the traditional scoring, which is to simply calculate the percent of the words within a text.

[RC 2.3.] (3) **As a communication scholar, I find the way outlets' leanings are measured by the composition of their readers at least questionable. I know it is a convenient shortcut, but again the question is whether the inferences drawn in the paper are really about the target audience, or the journalists' attitudes and moral predispositions (or the attitudes of the politicians quoted in the news articles). I encourage the authors to more thoroughly consider what kind of theoretical foundation their own arguments (or the cited previous applications of MFT to texts) have. A helpful source could be Benoit, K., Laver, M., & Mikhaylov, S. (2009). Treating words as data with error: Uncertainty in text statements of policy positions. American Journal of Political Science, 53(2), 495-513.**

While we appreciate the complexities of measuring partisan lean of media outlets, we justify the selection and categorization of the news outlets using a Pew Research Center study. Obviously other methods could have been used, but we feel that this is sufficient justification.

"We determined the political lean of each source by referencing (**Mitchell2014?**)'s article demonstrating the self-reported ideological consistency represented by the consumers of several news sources."

[RC 2.4.] (4) **The manuscript can be shortened by removing redundancies between study 1 and 2 descriptions. Also, if the text statistics such as readability are not used as controls or otherwise in the main analysis, why bother with them? While I appreciate the excellent documentation of the scraping,**

much of this can be moved to a supplement.

We have streamlined in method sections when it was repetitive referring back to previous studies. We did keep places in which analyses were different or otherwise it was necessary for clarity (given the other reviewers comments).

[RC 2.5.] **(5) I was really irritated by the use of "Experiment" 1 and 2 in the headings, when no experiments were conducted or at least reported. Please stick to study 1 and 2.**

The wording has been changed throughout the document.

[RC 2.6.] **(6) Reporting the ICC of the Null model with just outlet clustering would be nice and helpful to understand the data.**

These values have been added to each model table.

[RC 2.7.] **(7) Are we sure that linear models are appropriate for the weighted frequency outcomes?**

Each model was checked for assumptions of parametric models. Linearity, normality assumptions were met with some violations of homoscedasticity.

3. Reviewer #3

[RC 3.1.] **In this article, the authors provide a two-study investigation of the moral leanings of news sources using an approach based on the Moral Foundations Dictionary (MFD). This paper has a number of strengths, and I do believe that we need more work carefully validating dictionary-based approaches to studying moral content (and moral foundations theory) in textual data. I would also like to emphasize that this work is especially impressive being a Master's thesis. This being said, I also see a number of issues with this manuscript that, in my view, limit the validity of its conclusions. In short, I found myself wondering throughout the methods section(s) as to the extent to which the results observed in this manuscript are due to the analytic decisions that were made rather than due to issues with the MFD (or MFT). This concern is magnified when considering that the authors used a number of rather non-standard techniques compared to extant work. I'll highlight a few of the most salient issues below: My first concern is the most central to my overall evaluation of the manuscript. I noticed a large number of places where critical methodological details of the analyses conducted in this manuscript were completely left out or glossed over. I appreciate that the authors provided supplemental material containing analysis code, and I'm sure that I could dig through the code and figure out the answers to many of these questions, but it's critical that methodological choices that are central to the conclusions of the paper be provided in the manuscript itself. A couple of noticeable examples:**

• The details of the valence-weighting procedure applied to the moral foundations words was not clear to me, even after several readings. Was the valence calculated for the moral words themselves, or for the context surrounding the words? How were the z-scores applied (within i.e. within document, within source, overall)? These details are important, because the mean valence of the MFD words

varies across different MFD categories, and also because the valence of moral words is almost always contingent on their context (see e.g. Hopp et al., 2021). For example, the sanctity-assigned word "pure" could be be used to refer to "pure water" or to "pure evil."

We added an example table of the weighting calculation and rephrased this section completely for clarity. Unfortunately, as with many linguistic analyses that focus on individual words (as with the MFD), context is lost using these approaches.

In response to the question about analytic choices, we have added information about a MTMM study (study 1), as well as the exact same statistics using the mean percentages for comparison (study 2-3). We believe these additions will allow for comparison of different approaches versus the normal approach and the validity evidence for the MFD.

[RC 3.2.] **• The details of the MLM's are also unclear. For example, it seems like political leaning is treated as a categorical variable, but that isn't mentioned. It's also unclear how the random intercept for source was created. Again, I could be misunderstanding something, but if source wasn't nested within political leaning, then creating the random intercept for source would almost certainly wash out any effect that may be attributed to political leaning. I would suggest providing the equations for the models within the text to avoid these sorts of confusions.**

As described in the text:

"Owing to the lower number of sources analyzed herein, we elected to categorize the sources as either"liberal" and "conservative" in order to form a basis for comparison."

We would also note that the complete manuscript and code is available on our OSF page: `https://osf.io/5kpj7/`. We added the equation for the multilevel model directly to the text as well:

"Therefore, the models were calculated by: Weighted Score ~ Political Lean + (~1|Source)."

[RC 3.3.] **My second concern stems from the first. Given the lack of methodological clarity at several key points in the manuscript, I was somewhat surprised by the forcefulness of the authors' conclusions. To be clear, I say this as someone who is also quite skeptical of the validity of the MFD (for a more extensive treatment of why, see Hopp et al., 2021). Given that the authors used a customized version of the MFD, I don't personally see how evidence that the predictions didn't pan out is evidence that the MFD (much less MFT in general) is less valid.**

We added comparisons to the original scoring to demonstrate the results remain the same (study 2 and 3).

[RC 3.4.] **This is especially true given that a good deal of the work the authors reference just looks at word counts (not valence-weighted percentages) as indicators of moral-foundations-related differences across sources. As a final note, I would suggest that the authors go through the manuscript with an eye toward including everything that is needed for an interested reader to know the critical methodological details, but also toward removing things that are not germane to the questions that are the authors' central foci in the manuscript. For example, I would presume that the linguistic complexity analyses could be relegated to supplemental material to make space for more methodological discussion. In sum, although there is much to like about this manuscript, I believe that it needs quite a bit of re-engineering before it's ready to be published. I would again like to commend the authors on their efforts. I hope that my comments are helpful in continuing to refine this important work.**

We have moved information to the supplemental document, as suggested by several reviewers. We additionally add more information about the methodology for clarity.