Dear Dr. Evans,

Thank you for your comments on our manuscript, "Implicit Ideologies: Do Right-Wing Authoritarianism and Social Dominance Orientation Predict Implicit Attitudes?" We'd also like to thank the reviewers again for the careful attention they've given our manuscript. We feel that the MS is stronger now because of the very helpful comments from the review team. We noted and considered the two more brief reviews, but primarily, we followed your suggestion to focus our revisions on the two more substantive reviews (from reviewers 1 and 3). Their comments and our responses are summarized below, but we are also happy to elaborate further on any of the responses, if needed.

Reviewer 1
I like that you discussed implicit preferences by considering their limitations and critical issues. This is of extreme importance. However, I was left with the impression that it does not emerge why to care about implicit associations. In the paragraph "Implicit vs. Explicit Attitudes" (pp. 4-5), it would also be important to note why it is relevant to investigate implicit attitudes in addition to explicit ones. E.g., implicit measures are less susceptible to demand characteristics and can help measure responses that may be introspectively inaccessible. Further, even though implicit and explicit measures appear to assess the same underlying attitude in some research, they can still tap into more spontaneous evaluative reactions and more controlled responses. In other words, although the attitude might be the same, its expressions might differ.

We appreciate this feedback to clarify the reasoning behind the current study. The manuscript has been updated to note the important differences between implicit and explicit measures, with relevant citations. That is, we have highlighted the importance of capturing implicit attitudes in addition to explicit attitudes, due to the less conscious answering typical of implicit measures. These changes can be found in the section "Implicit Versus Explicit Attitudes" starting on p. 5 of the revised MS.

In the next section (pp. 5-6), you stated, "Given the findings relating to the IAT and political orientation, we might expect RWA and SDO to predict implicit attitudes in a similar way to that of explicit attitudes. However, research employing RWA and SDO to predict implicit attitudes is relatively scarce and offers mixed findings." The mixed results might be due to the heterogeneity of the attitudes examined. Thus, I do not feel it is a compelling reason to explore implicit attitudes. Perhaps it might be more effective to stress instead that much research has shown that RWA and SDO are significantly linked to various attitudes. If SDO and RWA are predictive of those attitudes, you can expect both to be related to the implicit attitudes of interest, which are closer and more relevant to the considered ideologies (which you noted in the abstract but not elsewhere in the RR).

This is an important point as this touches on the key research question underlying this proposal. In the section titled "How Do RWA and SDO Predict Implicit Attitudes?" (p. 6), we discuss how the previously discussed attitudes linked to RWA and SDO have less clear relationships when using implicit measures. At the end of this section, and within the "Present Study" section, it is highlighted that this remains an important research question to be addressed.

Another major issue I see in the RR's introduction is that it lacks a clear description of the research question and hypotheses. Indeed, from the "Present Study" section, it is hard to understand what the dependent variables are and how RWA and SDO are expected to relate to each of them (both in terms of direction and magnitude of the effect). For example, as you provided examples of IAT pairings that should be more relevant to RWA than SDO and vice versa, I find it puzzling that you did not speculate on differences in effects direction and size for each of them. Related to the dependent variables, I learned later in the RR that you plan to analyze 13 concepts (implicitly and explicitly). However, it is unclear to me what criteria you followed to select them, as for some pairings, it might not be evident why they are relevant (e.g., 1950/2050 IAT pairings). Further, what pairings are relevant to either SDO or RWA and which are relevant to both?

Initially, the full list of IAT pairs and expected relationships with RWA and SDO was not included in order to improve readability, but this ultimately did leave the hypotheses too vague, so we appreciate this feedback. In response, the full list of IAT pairs/items, their associated word pairs, and their hypothesized relationships to RWA and SDO are included in Table 1 (which can be found on pp. 9-11) for a concise, but full, summary. In this table, we also summarize the theoretical basis for each prediction, including citations to supporting research. Please note that clarity and concision, we have also removed one IAT pair because it was the only stimulus associated with "danger/safety" as opposed to "good/bad".

Although data collection and measurement details are described elsewhere, I believe this section should stand on its feet. For example, it should be clear what procedure was used to collect data (e.g., order of the measures, counterbalancing), data exclusion criteria (e.g., how to treat outliers), how variables will be computed (e.g., I guess explicit attitude will be the difference between the preference for x and y), and the sampling plan (which is briefly described in the Table under Study Design section) should be presented in more details. Further, it might also be helpful to understand the minimum effect size detectable given the available sample size (sensitivity power analysis).

Thank you for these suggestions. We agree that these methodological details should be included. We have updated the "Method and Materials" to add these details. Specifically, the "Design" section includes details of the data collection. Footnote 1 details the IAT exclusion criteria (there are no other exclusions). For each measure (including preference between x and y) we now report the scale format and anchors. We've also followed the suggestion to add a sensitivity power analysis (this can be found on p. 19 in the "Power Analysis" section).

Further, I believe the analysis plan should be enriched with relevant information, such as how the variables will be treated in the SEM model and the fit indices you will use to evaluate the goodness of fit. Further, do you plan to check whether the missing mechanism is as expected by design (in planned missingness design, missing data points are MCAR by definition; thus, you might use Little's MCAR test to check this on your data)?

Thanks for this suggestion. We've added sections on the modeling details starting on p. 14 (see the sections headed "Measurement Models" and "Path Models"). We also now report how we will assess model fit (we are planning to report dynamic fit indices as suggested by McNeish & Wolf, 2023).

We've also clarified the missingness assumptions. For a planned missingness design the data can be assumed to be at least missing at random (MAR), meaning that missingness is not related to unobserved characteristics of the data: "In certain settings, MAR is known to hold. These include planned missingness in which the missing data were never intended to be collected in the first place…" (Schafer & Graham, 2002, p. 152). Unbiased multiple imputation requires only MAR (not MCAR) (see, e.g. Enders, 2022). We've clarified this on p. 13 of the manuscript.

As I learned from the OSF documentation, data were collected from different countries (e.g., WEIRD and non-WEIRD countries). Thus, how do you plan to control the hierarchical structure of the data and measurement invariance, especially on SDO and RWA measures (there might be inconsistencies in the measurement models across cultures)?

This is an important concern. Because the majority of Project Implicit's respondents were from the US, we have decided to limit the data to participants from the U.S. in order to avoid concerns about measurement invariance across countries. We explain this logic on p. 12 ("Participants" section).

Minor comments:

P. 5, row 3: Please, add a reference to IAT.
P. 5, row 7: Remove "toward" from the text.
P. 5, row 5-7: From the description, it should emerge that the implicit association measured with the IAT is relative in nature.
P. 6, row 14: "Emphasize" needs the s.
Sometimes words are spelled in British, and some others in American.

Thanks, we've fixed these mistakes.

Reviewer #3

1A. The scientific validity of the research question(s).

As I understand it, there are two (or more) research questions here: the first concerns the relation between self-reported RWA / SDO and implicit attitudes about related topics. This question is justified: the existing evidence base has produced mixed findings.

Another question concerns whether implicit and explicit attitudinal measures are capturing the same underlying construct, and the validity of the IAT ("are implicit attitudes consistent with explicit?"). At this point, it is unclear to me how exactly these two questions will be evaluated, that is, which parts of the design will be used to draw conclusions about these questions. I do not fully understand how the validity of the IAT could be tested with this design – wouldn't differences between implicit and explicit attitudes be expected from an IAT perspective? I would *appreciate if the authors could describe their reasoning and expectations with regard to these questions in greater detail, and link them more clearly to the analyses that will be conducted.*

We agree that the logic for the secondary research question could be clearer in the manuscript, so we have updated that. Specifically, while explicit and implicit measures are expected to have some differences, they are viewed as generally being similar, "additive" methods of measuring underlying beliefs. That being said, observing similar response patterns using both explicit and implicit methods would support the idea that RWA and SDO could similarly predict explicit and implicit attitudes. Further, addressing a less central research question in the current study, a similar pattern of responses between implicit and explicit measures would also *support* the validity of the IAT in measuring the same underlying attitude as explicit measures. Of course, this design would not offer any definitive evidence, but would help add some information to the bigger picture.

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.
The authors propose the hypothesis that "RWA and SDO will predict implicit attitudes in line with explicit attitudes toward a range of relevant topics". In my opinion, this hypothesis is not sufficiently precise and contains multiple aspects that should be more clearly differentiated.

First, I think that the relation between RWA /SDO and the respective explicit attitudes should be tested in a separate hypothesis. It may be that, contrary to expectations, these relations do not emerge in the predicted fashion (although it appears unlikely).
Second, I would like the hypotheses to clarify which precise topics are expected to be related with RWA and SDO, and which topics are expected to be differentially related to RWA / SDO. That is, are there topics for whom you expect only an association with RWA and not SDO (and vice versa)? In the introduction, it is hinted that such differential relations exist, but then this is not further addressed.

As per the first comment in 1B, we agree that this could be broken up into two, more clearly testable and transparent hypotheses. That is, one hypothesis predicting that RWA and SDO will predict *explicit* attitudes in a way that is consistent with the ideology of each scale, and a second hypothesis predicting that RWA and SDO will predict *implicit* attitudes in the same way. As mentioned in our response to reviewer 1, the full, specific list of hypothesised relationships is now listed in Table 1.

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).
There are several aspects with regard to this criterion that require clarification.

First: Which participants will be included in the analysis, and why? Will everyone from this data set be included? As the data are already collected, the authors could provide much more specific information about the sample size and power considerations (beyond the rule of thumb mentioned in the Study Design table).

We have added much more detail about which participants will be included, and what the sample size for each analysis will be. These details can be found in the "Design" and "Participants" sections (p. 12 and 13, respectively) as well as in Table 1. We also report a sensitivity power analysis showing that we have power of at least .98 to detect our minimum effect size of interest in each test individually.

With regard to the analysis pipeline, the authors say that they will conduct SEM with FIML. Many questions remain open here: what will the precise SEM look like? I would like to see the specific paths that will be estimated, ideally in a graphical depiction. Which variables are estimated at the latent level? What does the measurement model look like?

We agree that these details ought to have been included. As we explain in the text, we will estimate separate measurement and path models (see pp. 12-15). Single-factor measurement models will be estimated separately for RWA and SDO, and estimated factor scores for each scales will be used in separate path models. One example path model is shown in Figure 1.

More generally, will one large SEM including RWA, SDO, implicit and explicit measures be estimated? Will there be separate models? How will model fit be evaluated? How will bad fit be dealt with? Which parameters are relevant for testing the hypotheses? Ideally, the authors would describe the SEM they plan to conduct in detail and describe which parts of the model test which hypotheses.

For each topic, we will estimate separate path models for RWA and SDO, which will include both implicit attitudes (one example is shown in Figure 1). We now also now state the hypothesis test(s) explicitly:

"For each model, the path coefficients between RWA/SDO and explicit ratings/D-scores will test our predictions" (p.16).

As noted above, our approach to assessing model fit is to use dynamic fit indices as proposed by McNeish and Wolf (2023). Even in the case where the measurement model fit is bad according to these indices, we plan to continue to the path-modeling part of the analysis pipeline. Another option would be to make empirically-based modifications to the measurement models to improve the fit (i.e., by using modification indices). However, we think it's important to maximize comparability to the previous literature, which has treated these scales as unidimensional and has typically computed simple composites. We do think that reporting and discussing model misfit (if it exists) will be an independent contribution to the literature and may encourage future researchers to rethink how these instruments are typically used.

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses

Although a lot of documentation about this data set appears to be available on OSF, I would appreciate if the authors could provide more information about the methodology and the measures in their manuscript. Especially with regard to the implicit and explicit attitudes that will be analyzed, it should be clearly stated which topics will be analyzed, and how these were *measured. Currently, the authors state examples of topics using "such as", which implies that not all topics that will be analyzed are mentioned.*

Also, could the authors describe the planned missingness design in a bit greater detail? I am not familiar with this approach. Why did participants in design A receive one, and those in design B two topics? Does this have any implications for the analyses (e.g., should the type of design be included as a random effect in the model)?

The full list of analyzed topics is now specified in the manuscript (Table 1). The differences between Design A and B were designed as a tradeoff between measurement breadth and depth. Participants in Design A only received one topic because they completed a total of 9 explicit measures on the one topic, while participants in Design B completed 1 explicit measure on two different topics. Therefore, for consistency, we have only included the one explicit measure that every participant completed (a relative liking/preference item), regardless of which Design they were assigned to.

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).
As far as I understand, the authors did not yet include any outcome-neutral conditions. In the context of the SEM, I think this could pertain to considerations of model fit (as mentioned above), and how bad fit will be dealt with.

Thanks for this suggestion. We will include item distributions (histograms and descriptive statistics) for all variables in the model in the Supplemental Material. As noted above, we will proceed with path modeling even if there is bad fit for the measurement portion of the model. However, we do think the reviewer is right that it is important to have a positive control. An obvious sanity-check is that SDO/RWA scores ought to correlate with self-reported political ideology (i.e., liberal-conservative). We will report these correlations in the main text (see the "Positive Controls" section on p. 16).