Dear Dr. Evans,

Thanks to you and the review team for your careful reading of our MS. We are grateful to the two reviewers for their insightful comments. In response to these comments, we've substantially restructured the Results section and have added to the General Discussion. Point-by-point responses to each comment are below. Thank you again for your comments and for the opportunity to revise the MS.

We'd also like to point out one other change to the MS in addition to those in response to reviewer comments—in the process of checking the code and results before resubmission, we discovered that some of the counts reported in the Stage 1 version for the number of participants who completed SDO and RWA scales differ somewhat from the final dataset. This is because:

1. The Stage 1 analyses were done on a blinded version of the data, and the Project Implicit variable we used to determine whether the participant completed the scale does not correspond exactly to whether the CFA models can estimate a score for that individual.
2. For the overall (measurement model) Ns, in the Stage 1 analyses we erroneously didn't exclude non-U.S. participants from those counts.

For now, we have updated all counts throughout the MS to be consistent with the final data, but this does involve minor changes to some of the sections that were originally approved in the Stage 1 acceptance. Please let us know if you would like us to handle this a different way.

Reviewer 1

According to the method, you tested 24 models (12 tasks for the two independent variables). This is also evident from the R script where you estimated the models using a for loop. Therefore, it is unclear why the fit indices are reported for only two models. I strongly recommend clarifying this point.

This is because we estimated the measurement models (for RWA and SDO) separately from the path models. For the path models, they are fully-saturated (every path estimated) with no measurement portion, so fit will be perfect. We report fit for the measurement models on p. 19-20 in the first paragraphs of the Results section.

We understand why the reviewer had this question, because in retrospect the Results section was badly structured; it combined the measurement and path model results in a confusing way. We've restructured it to follow the same progression as the "Analysis Pipeline" section above, so we now first discuss the measurement models and fit statistics, then the path modeling results.

From the script, it is also evident that you tested the measurement model for RWA and SDO. For clarity, I would report the results of these models, as they provide important details on the quality of the measurements, which could be relevant for contextualizing the results of the structural models. For example, the lower predictiveness of SDO might be due to measurement issues of the construct.

As we note in the previous response, the measurement model results (i.e., fit statistics) are shown on p. 19-20 of the revised MS.

Additionally, you used dynamic cut-offs. As reported in the cited paper, this method can be applied to measurement models in the SEM framework but not to structural models as you did. Moreover, the method is uncommon, and reading the results without contextual information makes understanding difficult. For instance, what are the implications of choosing a cut-off with a low level of misspecification versus a high level? I strongly suggest providing more information on this method to improve reader comprehension.

The dynamic fit cut-offs are only used to evaluate the measurement models. We've moved up the table showing these cut-offs and incorporated the interpretation into the paragraph on p. 20 discussing model fit. We've also added a note to what is now Table 3 to help readers interpret the thresholds. Finally, we realized that our previous interpretation of the DFIs was not right, and that they (like the traditional fit statistics) point to bad model fit. We now note this explicitly (see p. 20).

To further refine the comprehensibility of your results, I recommend incorporating the B-H Critical Value present in Table 6 into Table 3.

In response to this comment and a similar comment from Reviewer 2, we now focus on the (preregistered) B-H critical thresholds. So, we now show paths that are significant according to the B-H threshold in bold in Table 4. The critical values and test p-values are shown in Table S1 in the Supplemental Material.

This comment bridges both the method and results sections. Therefore, I consider it a secondary addition, although it is highly relevant. The question I asked myself while reading the results is: how do implicit and explicit measures correlate for the same pairs of words? This could provide further details on whether explicit and implicit measures capture the same or a similar latent construct.

We have added these data to the General Discussion (p. 26):

"In general, implicit and explicit attitudes were moderately correlated (average *r* = .29), though the size of the relationship varied widely, from *r* = .49 ("socialism/capitalism") to *r* = .06 ("equal/unequal"). Correlations for each pair of implicit and explicit measures are shown in Table S2 in the Supplemental material."

Reviewer 2 (Luisa Liekefett)

Although the authors did not specify outcome-neutral criteria necessary for testing their hypotheses, the model fit according to the dynamic fit indices seems adequate. However, the absolute fit indices did not indicate adequate fit for the measurement models. Perhaps the authors could elaborate on whether this is cause for concern or not. I am not very much familiar with dynamic fit indices, and some more details on this topic could be helpful for the reader.

As we’ve noted above, we added some details to help the readers interpret the dynamic fit indices. However, the correct interpretation is of the DFIs is that the measurement models do not fit well according to either the DFIs or the standard criteria. Our Stage 1 analysis plan indicated that we would move ahead with analyses regardless of fit. We now discuss the sub-par fit for both models, and what it means, on p. 29.

Yes, the authors adhered closely to the registered study procedures. The only deviation I could find was that the authors focused their discussion on results with p-values that remained below the .05 cut-off, instead of the results that were deemed significant by the B-H procedure. This seems like a reasonable decision to me. However, perhaps the authors could briefly discuss if the overall conclusion would have been any different had one sticked precisely to the planned B-H procedure.

In combination with Reviewer 1’s comments about better integrating the B-H procedure results, this comment made us re-evaluate our approach of focusing on the $p < .05$ results rather than the registered B-H tests. Since the difference between the two are small (the B-H procedure deems two results with $p > .05$ significant) for simplicity and consistency with the planned analysis we now focus our interpretation on the tests deemed significant by the B-H procedure. However, readers can still consult the 95% CIs in Table 4 to see which results are significant at $p < .05$.

The authors conducted one exploratory analysis that is mentioned in the discussion. In this analysis, the authors examined whether relations between RWA/SDO and the respective implicit/explicit attitudes resemble relations between general political conservatism and these implicit/explicit attitudes. Although this analysis appears reasonable, it does leave me a bit unsure as to what this means for the interpretation of the main results. To what extent are the observed links specific to RWA and SDO, or result from shared covariance with conservatism more generally?

We think this is a great point, but this is difficult to test using the data we have. We do not think that the right approach is to statistically control for overall ideology, as it is so closely conceptually related to both SDO and RWA. Rather, we suggest that future research measure both SDO and RWA in the same individuals, to examine the effects of one controlling for the other. We have added this to the “Limitations and Future Directions” section of the General Discussion.

Do you have any idea as to why some associations did not turn out as expected? I found it surprising that the anarchy/hierarchy pair, for instance, was not associated with SDO, neither implicitly nor explicitly. You already mention in the discussion that the relationships appear to be sensitive to the specific wording of a construct. What could these differences in wording be, and how could they explain the observed findings?

We’ve elaborated on one possibility in the General Discussion on p. 28:

“[It may be] that some of the word pairs tested represent constructs that are unfamiliar to participants or that they have not spent much time thinking about. If this were the case, it would be expected to lower the reliability of IAT scores (Cummins, Hussey, & Spruyt, 2022), which in turn would reduce power to detect any effects. For explicit measures, where conscious deliberation has room to operate (Wilson et al., 2000), unfamiliarity may be less problematic.

What are the broader (theoretical) implications of these findings for the literature on implicit attitudes, the validity of the IAT, and/or the measurement of RWA/SDO?

Thanks for raising this important point. We've added a new General Discussion section called "Theoretical and Methodological Implications" (see p. 29) to discuss some high-level take-aways from the results.

Perhaps I missed something, but I would be interested in the covariances between the implicit and explicit measures. In the introduction, you mention that one would not expect implicit and explicit attitudes to correlate perfectly, and that they are usually correlate around r = 0.30. Do you observe converging findings? Do these covariances support the conclusion that both measure the same underlying construct?

Please see the last point in response to Reviewer 1 above.