Editor’s comments

I have now received two very detailed and constructive evaluations of your submission. As you will see, the reviews are broadly very positive while also raising a range of points for further consideration, including a range of core issues relating to study design (including suitability of controls and potential carry-over effects), calls for additional methodological detail, and clarification of the analysis plan and key aspects of the study materials.

Concerning the point raised in Matt Williams' review regarding preregistration of the pilot studies, I agree that reporting of these details in the main text is less crucial given the purpose of this preliminary research; however it is always good practice to mention preregistration (and changes from protocol) where applicable, so please do mention in the main text of the manuscript where specific pilot studies were preregistered on AsPredicted. In addition, please document all relevant deviations from those protocols in the Supplementary Information file.

I hope you will find these reviews helpful in further strengthening your (already impressive) proposal and look forward to receiving your revised manuscript in due course.

Response:

We thank you and the reviewers for these detailed and helpful comments. We greatly appreciate the constructive and thorough peer review process. We have done our best to address all of the comments below. We have made several major changes, particularly with regard to the sampling plan:

We no longer plan to recruit participants on Prolific. Instead, we plan to use the non-commercial SoSci Survey Panel. This has three major advantages: First, participants are genuinely motivated to participate in surveys, which should increase data quality and compliance (see Matt Williams´ comments). Second, only about 2800 German-speaking participants are currently active on Prolific; this number is reduced to about 2000 if one filters out those with a mental health condition. Thus, if it came to the final stage of the sequential design, the number of available participants on Prolific could be tight. In comparison, there are currently about 80,000 active panelists (from German speaking areas) registered in the SoSci Panel. Additionally, the panel is free of charge for non-commercial research of high quality.

We have already contacted the SoSci Panel with our proposal and they confirmed that our study could be implemented with them. They offered to collect 1,000 participants for the first *stage, and, if necessary, another 800 for the second stage of the sequential design. For this reason, we have adjusted the sequential design to have two stages (instead of three).*

Before starting data collection, the SoSci Panel will conduct a peer review of the questionnaire. This is to ensure that the questionnaire fulfills their criteria, which mainly relate to the length, visual design and user-friendliness of the survey. This review process will take approximately 4-6 weeks, and may lead to some (minor) changes in the questionnaire. As soon as we have received feedback from the SoSci Panel, we will inform you about any changes that may have been made in the process.

Note that we have already received confirmation that the basic features of the survey (i.e., number of participants, repeated-measures design, overall topic and length of studies, and the nature of the manipulations) are feasible.

Further information on the SoSci Panel can be found here:

Leiner, Dominik J. (2016). Our research's breadth lives on convenience samples. A case study of the online respondent pool "SoSci Panel". Studies in Communication | Media (SCM), 5(4), 367–396. doi:10.5771/2192-4007-2016-4-367

Leiner, Dominik J. (2012). SoSci Panel: The Noncommercial Online Access Panel. Poster presented at the GOR 2012, 6th March, Mannheim. Available from https://www.soscipanel.de/download/SoSciPanel.GOR2012.pdf.

In addition, we have prepared prototypes of the questionnaires that can be tested (and commented on) by you and the reviewers in English (note that we used an automatic translation, so some wording might not be ideal). The goal of this was to give you an impression of what the study and the manipulations would look like. Some features of the actual survey do not work in the prototype, because they require participants' unique identification number: the depression/suicidality screening, and the replacement of "worrytopic" with the topic participants actually selected in the T2 questionnaire. We also adjusted the manipulations to contain only one cycle of repetitions to save time in testing.

The prototpyes can be found here:

T1: https://www.soscisurvey.de/regreport/?act=2ttwhD6lxBaQk70DtFA1JTr2

T2: https://www.soscisurvey.de/regreport/?act=MYwjyVcPtt8tvaPq6fmkbuhq

Reviewer 1

Reviewer 1.1

The submitted registered report addresses an important theoretical question related to one of the potential antecedents of conspiracy beliefs: ruminative thinking. Although there is prior empirical evidence assessing the association with different types of thinking (e.g., analytical thinking) and cognitive biases, I consider the present research informative insofar as it delves into the dysfunctional cognitive and emotional elements arguably characterizing the thinking processes often associated to conspiracy beliefs.

I briefly provided comments regarding the theoretical introduction, the already conducted studies, and the interpretation of the available results, and I then focused on the proposed follow-up study that you plan to conduct.

Response:

We thank the reviewer for their helpful and encouraging comments. We did our best to address them all below.

Reviewer 1.2

Introduction

I appreciate the use of a state-of-the-art definition of conspiracy beliefs, deviating from traditional definitions that tailored the phenomenon to fake or implausible conspiracies.

The explanation of the conceptual association between rumination and conspiracy beliefs through negative affect, negative cognitive biases, and persecutory delusions is generally convincing. However, there are a few points regarding the 2 latter issues that could be clearer.

Regarding negative bias thinking, I think the argumentation could benefit from explaining other biases associated with conspiracy beliefs (e.g., agency perception; Douglas et al., 2016; catastrophizing; Green & Douglas, 2018), in order to establish a clearer link with those biases that rumination favours.

Response:

We thank the reviewer for this suggestion. We agree that mentioning other biases related to conspiracy beliefs can strengthen the argument. We have included a paragraph that refers to those biases on p. 8:

"In line with this, recent research has shown that conspiracy beliefs are related to a general suspicious processing style, that is, an intuitive tendency to perceive negative intentionality and secrecy in both conspiracy-related and -unrelated events (Frenken & Imhoff, 2022). Further, conspiracy beliefs are associated with several other thinking biases, such as the tendency to attribute agency and intentionality to inanimate objects (Douglas et al., 2016). An anxious attachment style, which entails an exaggerated perception of threat and a negatively biased view of others, has also been found to predict conspiracy beliefs (Green & Douglas, 2018). These findings show that styles of thinking that share properties with rumination contribute to the formation of conspiracy beliefs."

Reviewer 1.3

As for persecutory delusions, although I find the argumentation compelling, explicitly stating that "both [persecutory delusions and conspiracy beliefs] share the conviction that harm is going to occur" might not be totally accurate. It is not always the case that specific conspiracy beliefs are anticipatory of harm, but rather descriptive narratives of harmful actions in the past (e.g., 9/11 conspiracy theories). Of course, this can lead people to distrust the malevolent group responsible for these past actions and engage into new conspiracy beliefs about their future behavior, but this would not apply to the conspiracy belief that originated after the original threatening event.

Response:

We thank the reviewer for this comment. We agree that not all conspiracy beliefs entail a direct anticipation of harm. Although, as the reviewer describes, even those that are more descriptive of past events would likely go along with the concern that something similar could happen again in the future, and would thus be closely linked to an anticipation of harm. Nevertheless, we have restricted this statement. It now reads: "Nonetheless, similar to persecutory delusions, conspiracy beliefs entail the conviction that harm is going to occur (or already has occurred), and that a powerful agent (persecutor or group of conspirators) will cause (has caused) harm (Freeman, 2007)."

In addition, we describe in a footnote on p. 9:

"Please note that not all conspiracy beliefs contain an anticipation of harm. Some are conspiratorial interpretations of ongoing or past events (e.g., 9/11 conspiracy beliefs)."

Reviewer 1.4

Pilot 1

I would appreciate if you could justify the use of two rumination measures. Do they measure exactly the same? If not, how do they differ? This could be helpful for the reader, considering that you point out earlier the relevance of distinguishing between different elements of rumination (brooding vs. reflection).

Response:

We thank the reviewer for this comment. Both rumination scales measure broadly the same construct (i.e., repetitive negative thinking), although with a slightly different emphasis. We have included a paragraph that describes this on p. 10:

"Both rumination scales measure the broad tendency to engage in repetitive negative thinking. The PTQ focusses on the general characteristics of the thinking process (i.e., whether it is repetitive, unproductive, and/or intrusive), whereas the rumination subscale of the HFERST refers specifically to distressing events and ruminating about the causes of one's negative emotions."

Reviewer 1.5

I would warn the reader that, beyond this study being a correlational study, your statistical power is also not that high for the detection of half of the correlations you observed (the minimum correlation you can detect with your sample size is $r = .2153$, with 90% power and alpha .05). So these preliminary results should be taken cautiously.

Response:

We appreciate this comment. It is true that the power for the test would not have been high for small correlations. However, we believe that this does not make our result less trustworthy. As Mayo (2022) *states, "if POW(μ') [the power of the test to detect μ'] is high then a just significant result is poor evidence that μ > μ'; while if POW(μ') is low it's good evidence that μ > μ' (provided assumptions for these claims* hold approximately). *(...)* *Because it is improbable to get as low a P-value as we did (or lower), were μ as small as μ'–i.e., because POW(μ') is low–it is an indication we're in a world where population mean μ is greater than μ'" (Mayo, 2022).*

Nevertheless, assuming that the true correlation is indeed as small as the smaller correlations we observed, future research attempting to replicate the effect would do well to use larger samples than we did in Pilot Study 1 to achieve acceptable Type II error rates. We mention this in a footnote on p. X: "Note that, given our sample size, the achieved power for some of these correlations (assuming that they reflect the true correlation) was not that high (e.g., we would have had a power of 66% for a correlation of .16 with alpha = 0.05). Future research attempting to replicate these correlations should ideally use larger samples."

Mayo (2022): https://errorstatistics.com/2022/05/02/do-underpowered-tests-exaggerate-population-effects/

Reviewer 1.6

Pilots 2a and 2b

In the explanation of your design, I would change "participants were randomly assigned to a rumination AND a control condition" to "participants were randomly assigned to a rumination OR a control condition" to make it clearer that your design is between-subjects.

Response:

We thank the reviewer for spotting this. We have changed it to OR.

Reviewer 1.7

As for the manipulation in these two pilot studies, I had some comments (i.e., individual relevance of the topic for engaging in rumination, variability in how much people wrote and how much spent ruminating about it). However, you partially addressed them in Pilot 3, and certainly in your proposed study, so I will not comment further.

Response:

We are glad to hear that the reviewer sees their concerns addressed in the proposed study.

Reviewer 1.8

Pilot 3

Manipulation

It is a shame that the effect was not there, however, despite the manipulation checks being optimistic, I still wondered whether this type of manipulation can induce rumination through an online survey setting. One concern is the time, which might not be sufficient for people to engage in actual intense rumination.

A second issue is the one on which you base the justification of your follow-up study, namely, that the manipulation could have induced a mix of brooding and reflection. In the end, some of the manipulation checks do not allow one to distinguish between brooding and reflecting on a negative event. One can analytically approach an event with negative consequences, reflect on it, and experience negative affect due to the valence of the event, but not necessarily due to the type of thinking process.

Yet, I believe that the new brooding condition you proposed in the registration of your follow-up study can help to counter this issue, by focusing participants on the negative emotional aspects associated with the societal topic, and importantly, a reiteration of those thoughts. This will hopefully do the trick! Otherwise, you may need to think of other study settings (perhaps in the lab or an ambulatory assessment?) to provide participants a longer time window to experience this ruminative, recursive thinking.

Interpretation of available findings

I very much appreciated the clarity with which you discussed the set of mixed results and their potential explanations. It is never easy to describe and interpret such a puzzling set of findings!

Response:

We thank the reviewer for these encouraging comments and suggestions. We are glad to hear that they agree with our reasoning about the distinction between brooding and reflection, and the respective changes made in the proposed study. We also think that future studies should make use of additional study settings and will discuss potential limitations of the online setting in the discussion.

Reviewer 1. 9

I agree with you that the prediction regarding reflection is less clear. I think it is important to mention the temporal dimension in which reflection happens within the process of internalization of conspiracy beliefs. While reflection can protect against engaging in conspiratorial thinking when a threatening event occurs, for conspiracy believers, reflection could contribute to strengthening the justifications supporting the conspiracy narrative (see van Prooijen et al., 2020).

Van Prooijen, J.-W., Klein, O., & Milošević Đorđević, J. (2020). Social-cognitive processes underlying belief in conspiracy theories. In M. Butter & P. Knight (Eds.), Handbook of Conspiracy Theories (pp. 168-180). Oxon, UK: Routledge.

Response:

We thank the reviewer for bringing this helpful reference to our awareness. We have included this idea in the Supplement where we elaborate on the relation between reflection and conspiracy beliefs:

"However, we argue that the effect of reflection on conspiracy beliefs should depend on contextual factors, for example the extent to which one is already invested in the idea of a conspiracy, and motivated to bolster and justify this intuitive belief. As van Prooijen et al. (2020) describe, deliberate "System 2" reasoning processes about conspiracy suspicions can have two distinct consequences: they may lead some people to 'unbelieve' a conspiracy (which is consistent with the negative correlation typically found between conspiracy beliefs and analytic thinking), but they may also serve to reinforce preexisting intuitions that 'something fishy is going on' in a process of motivated reasoning. It may be that reflection plays a protective role when it occurs early in the thinking process (as soon as a triggering event is encountered), but that reflection reinforces conspiracy beliefs if an initial suspicion has already formed and is subsequently reflected upon (in particular, when the individual is already invested in the idea of a conspiracy in the sense that 'it simply feels true', and engages in motivated reasoning to support this suspicion, van Prooijen et al., 2020)."

Reviewer 1.10

Registered Report of Follow-up Study

As I said, I think this study could address some of the issues of Pilot 3.

I have some concerns about potential carry-over effects between T1 and T2, especially considering the short time between the two assessment time points (minimum of 24h) and the emphasis made in introducing the concept and the definition of conspiracy theories within the study. I believe that these two features could artificially increase conspiracy beliefs in T2, and reduce the strength of your manipulation. In the end, participants might already anticipate that they have to think about the "societal topics" in conspiratorial terms if they have read about it 24h ago.

One suggestion would be to increase the distance between T1 and T2 (maybe a week), although I understand that this may come at the expense of higher dropout rates and therefore higher data collection costs. Alternatively, you could consider masking the target societal topics (and related-conspiracy belief measures) among 4-6 other unrelated topics and items measuring non-conspiracy beliefs. In a similar line, I would suggest completely omitting the detailed preamble/definition of conspiracy theories included in the study materials (i.e., "Conspiracy theories are often discussed in the media. A conspiracy means that influential people join together in secret to pursue a common goal…"). I do not clearly see the benefit of including this paragraph (maybe I missed the point). However, I see its potential risk in framing participants' mindset for the rest of the study (including T2) regarding the identification of conspiracy explanations in subsequent material (especially if only 24h pass between T1 and T2).

Response:

We thank the reviewer for these detailed comments. An advantage of switching from Prolific to the SoSci Survey Panel is that we can now increase the time between T1 and T2 without the risk of inflating data collection costs. We therefore plan a distance of 5-10 days between the two measurements. As you suggested, this should reduce the risk of carry-over effects.

We also thought about including distractors in T1. However, we would prefer to keep the survey as short as possible in order to not strain participants more than necessary and to not take up more of their time than needed. We also think that the depression and suicidality scales we include at the end of T1 for screening purposes might function like a distractor, but we see the time between T1 and T2 as the main factor that should help prevent carry-over effects.

With regard to the preamble/definition of conspiracy theories: Our goal was to prevent participants from quickly rejecting the conspiracy items because they want to distance themselves from 'absurd' conspiracy theories - similar to how some people would say defensively "but I am not a racist!". However, we do see the point that this text could influence participants' mindset for the rest of the study in ways that are difficult to predict. This might be a bigger threat to the validity of our findings than some people hastily rejecting the items. For this reason, we have decided to omit the preamble.

Reviewer 1.11

I very much like the distinction between the new “Brooding” and “Reflection” conditions. However, I think that some features should be kept constant across both conditions to avoid any confound due to the current differences, i.e., in the number of questions (7 vs. 4) and in their level of requested justifications (i.e. Why does this concern make you feel so bad?). Regarding this issue, you could for example substitute Q2 in the “Reflection” condition for: Which argument do you find particularly compelling in favour of this explanation being true? Why do you find this argument compelling?/ Which argument do you find particularly compelling against this explanation being true? Why do you find this argument compelling?

Response:

We thank the reviewer for this helpful suggestion. We have adjusted the reflection manipulation to be more similar to the brooding manipulation with regard to the number of questions and the overall structure of the questions.

It now consists of the following questions:

1. *What could be possible explanations for X? Please take a moment to think about this before writing down the possible explanations.*
2. *Which of these possible explanations do you think is the most plausible?*
3. *What speaks for or against this explanation actually being true?*
4. *What is a particularly compelling argument for this explanation being true?*
5. *What is a particularly compelling argument against this explanation being true?*
6. *Now that you have thought about this, please make a final judgement: How plausible do you think it is that this explanation is actually true?*
7. *What could influence your judgment one way or the other?*

Repetitions (at least one cycle, then until 5 minutes have passed):

1. *What could be another explanation for X that you think is plausible?*
2. *Repetition of questions 3-7*

For comparison, the brooding manipulation looks like this:

1. *What concerns do you have about X? Please take a moment to think about this before writing down your concerns.*
2. *Which of these concerns makes you feel particularly bad?*

3. *Why does this concern make you feel so bad?*
4. *How do you feel as you think about this concern? Please describe these feelings in as much detail as possible.*
5. *Which of these feelings do you find most uncomfortable?*
6. *Why is this feeling the most uncomfortable for you?*
7. *What would happen to you if you felt such feelings very intensely for a long time?*

Repetitions (until 5 minutes have passed; at least one cycle of repetitions):

1. *What other concern about X makes you feel particularly bad?*
2. *(questions 3-7 as above)*

Reviewer 1.12

The instructions for the Manipulation Checks section use weird and complicated wording (i.e., think about the 5min before we asked you the questions about the conspiracies). I would simplify it with the suggested wording below:

"During the 5 minutes, we gave you to think about X, to what extent have you…

· Had depressing thoughts about X.

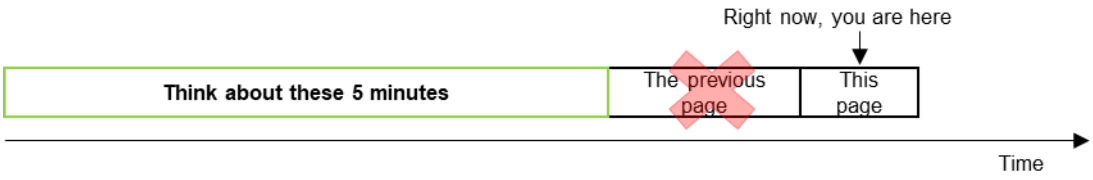
· Etc."

Response:

Thank you for this comment. We agree that the previous wording was somewhat complicated. We wanted to make sure that participants know very clearly what the MC items do and do not refer to: For the brooding and reflection condition, they should refer to the time spent with the manipulations. For the control group, they should refer to whatever participants did in the 5 minutes before answering the DV. For all conditions, it is important that participants do not include the time they spent thinking about the DV items in their evaluations of the MCs (to keep the measurement of the MC independent from the DV).

We have attempted to find a more intuitive way to introduce the MC. We now use the following introduction for all conditions:

"When answering the following questions, think about the 5 minutes before you answered the previous page of the questionnaire:



Reviewer 1.13

I also appreciated the detailed justification of the SESOI and I would not find a more compelling justification for a different effect size.

I think the detailed pattern of results you expect for the Manipulation Checks would be more easily summarized in a figure (bar plot?), highlighting the ones you finally use for your stopping rules.

Response:

We thank the reviewer for this idea. We have included the plot on p. 29.

Reviewer 1.14

Analysis Plan

I expected the "Reflection" condition to be included in the analysis plan, considering the justification given to interpret the mixed results of Pilots 1-3. I understand that brooding is the main condition of interest, as this should conceptually account for the effect of rumination on conspiracy beliefs. However, it is important to discard that reflection has a similar effect on conspiracy beliefs (implausible, but possible). Thus, I would like to see this included as part of the plan for the main analyses, which would entail changing your Welch t-test for an ANOVA framework with the 3 conditions as predictor of the difference in conspiracy beliefs between T1 and T2.

Response:

We have thought thoroughly about this suggestion. It is true that the mixed findings of the previous studies led to the distinction between reflection and brooding. If our goal were to only explain the pattern of results of Pilot Study 3, we should definitely have strived for a design that would allow us to include the reflection condition in our planned comparisons. For the outcome of such an endeavor to be informative, we would have to make sure that other changes in the design were minimal to avoid potential confounders. However, our main goal is to test whether a specific type of rumination, brooding, increases conspiracy beliefs.

This goal led to several changes in the design of the registered study compared to Pilot Study 3 (e.g., focus on within-person effects, the 5 minute and at least 1 repetition criterion in the experimental conditions). So, we would be hesitant that the pattern of findings from the registered study could be used to evaluate our putative explanation for the results of Pilot Study 3: Compared to the test of our main hypothesis (the effect of brooding), the outcome of this test would be very preliminary and tentative. For this reason, we thought it more transparent and adequate to use the findings related to the reflection condition in an exploratory fashion. Please note that we think not collecting any data on reflection would be a lost opportunity, exactly because of the many unknowns. We will try to use the data to inform future studies.

Reviewer 1.15

As for the sequential approach, it is unclear whether you are using corrected or uncorrected effect sizes for your stopping criteria and your equivalence tests. It is recommended to correct for bias when stopping earlier (in 1/3 or 2/3). This should be specified in your preregistration.

Response:

We are currently planning to use uncorrected effect sizes. Since the whole argumentation for our SESOI is based on Cohen's d and we cannot (yet) make any theoretically well-justified prediction for the impact of the conditions on variance, we would prefer to stick to this for our confirmatory test. Nevertheless, we will conduct robustness checks using Hedge's d and Glass's delta and will discuss if and how these corrected measures would have altered the conclusions.

Note that the difference between Hedge's g and Cohen's d is typically only noticeable in small samples (Lakens, 2022). Further, if our manipulations were to increase variance (in which case, Glass' Delta would be recommended), then Cohen's d would tend to underestimate the true population effect (Morris, 2008), thus simply resulting in a slightly more conservative test.

Hope you find some of these comments helpful, and I wish you the best of luck for this next study!

Response:

We thank the reviewer very much for taking the time and energy to help us improve this research. We greatly appreciate their comments.

Reviewer 2

Reviewer 2.1

Thank you for the opportunity to review this Stage 1 RR manuscript. I have organised my review according to the five criteria for PCI Stage 1 RRS, with two extra sections at the end. Overall I thought this was an extremely impressive manuscript, which displays a strong knowledge of the relevant literature, with appropriate and deeply considered methodological choices. My review here is quite long, but that's more of a reflection of the fact that I was very satisfied with the broad strokes of this manuscript, and focused on relatively specific details in my review. I am confident that the issues I've raised below are things the authors could feasibly address, so am happy to recommend R&R.

Response:

We thank Matt Williams for these encouraging words, and greatly appreciate how thoroughly they reviewed our manuscript. Their comments have been of great value to us.

Reviewer 2.2

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

I agree that the key research question (as specified in the manuscript title) is valid.

1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

The manuscript provides a coherent and sensible justification for the proposed hypothesis that "brooding about distressing societal issues increases conspiracy beliefs". The hypothesis seems sufficiently plausible to justify testing it.

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

In general I think that the planned methods and analyses are very sound, feasible, and rigorous. I was impressed especially by the effort that the authors put into planning the manipulations and the equivalence tests. I nevertheless have some minor/specific points to raise.

Response:

We thank the reviewer for their positive evaluation. We have done our best to address their comments below.

Reviewer 2.3

The paragraph at the top of page 19 discusses the fact that different participants will be ruminating about and answering conspiracy beliefs questions about different social topics. I agree that the use of randomisation means that this isn't a plausible confounding variable. That said, have the authors considered applying an analysis method that involves a random intercept for social issue? It's possible that accounting for variance due to issue might increase power, and this could also facilitate generalisations *across stimuli/issues* (see Judd et al., 2012; Yarkoni, 2022). This would obviously complicate the analysis plan somewhat so I consider this an *optional* suggestion, but I'd be interested to know the authors' thoughts.

Response:

Thanks for this important comment. We agree that, ideally, a mixed model with a random effect should be used. However, some features of our design complicate this procedure, which is why we would prefer to stick to the conventional t-test for our main analysis.

First, the number of levels of the grouping factor "societal issue" is quite low (we have only six different issues), which could result in reduced power of the mixed model (Singmann & Kellen, 2019). It is recommended that each level of the grouping factor should have at least 20 levels, otherwise power might be too low (Singmann & Kellen, 2019). In addition, it seems possible that some issues (especially the more right-wing oriented ones like "restrictions of freedom of expression") will be selected by only few participants, resulting in few cases in some of these levels.

In addition, we did not find any literature on how a mixed model could be combined with the equivalence and minimum effect test, which implications this model would have for the smallest effect size of interest, and how it could be incorporated into the sequential design (especially its power analysis and calculation of alpha and beta spending functions).

For these reasons, we would prefer to stick to the simple t-test for the confirmatory analysis, but will calculate and discuss a mixed model as a robustness check.

Reviewer 2.4

I can understand why the authors have used a no-distraction control condition, but it does strike me that this means that participants in the reflection and brooding conditions will spend quite a bit longer on the study than those in the control condition. Consequently, the study will implicitly manipulate payrate (per hour) along with rumination. The participants in the reflection and brooding conditions will also spend this extra time answering questions that are *open-ended* and *repetitive* – a task which Prolific participants tend not to enjoy (going by what I see them saying on the Prolific subreddit!) Consequently, the manipulation may affect irritation levels, which at least in theory could thus provide an alternative explanation for an apparent effect (e.g., maybe the "real" effect could be assignment to reflection or brooding conditions >- longer study -> lower payrate -> irritation -> belief in conspiracy theories). I don't think this is an especially strong or plausible threat to the validity of the findings, but it may be worth considering as a minor limitation.

Relatedly, from an ethical perspective, you might also want to consider whether it would be appropriate to pay participants in the more time-consuming conditions a bonus to compensate them for the extra time, or pay everyone a reward commensurate with the time of the longest condition.

It might be useful to build in an exclusion condition based on Prolific ID to check for inadvertent duplicate responses (this can happen when a participant has a problem with the survey webpage or something).

Response:

Thanks for paying attention to these points. The advantage of switching from Prolific to the SoSci Survey Panel (see our response to the Editor's comments above) is that participants no longer take part for the money, but because they are genuinely interested in scientific surveys. This already solves the issues with payment described above. Nevertheless, we would like to note that, had we used Prolific, we would have paid everyone the same amount (as if they had been assigned to the longest condition).

Reviewer 2.5

Re. statistical power, the manuscript says that "The suggested design simultaneously controls the Type I and Type II error rates of both the original null hypothesis test (which determines whether there was a significant effect on conspiracy beliefs) and the equivalence test (which determines whether values above d = 0.20 can be rejected): In the equivalence test, the original null and alternative hypotheses become reversed, so that the Type II error from the original null hypothesis test becomes the Type I error from the equivalence test, and vice versa (Lakens et al., 2021)." This would make sense if the authors planned only a conventional “nil” NHST plus an equivalence test, but as far as I can tell it doesn’t take into account the planned minimum effect tests. I.e., if the true d is 0.20, the power of the minimum effect test (where d = 0.2 is the “null” hypothesis) would be equal to a = 0.05, not 90%. Therefore, a power analysis for the minimum effect tests would need to hypothesise a true effect size greater than the SESOI of 0.20. Depending on what hypothesised effect size is chosen this could have substantial implications for the required sample size.

Response:

We thank the reviewer for this important comment. Indeed, we had not considered the power of the minimum effect test before. It is important to note (and we clarified this in the manuscript) that we have powered the study to detect the SESOI in a conventional nil NHST, and not for the equivalence and minimum effect test (this is in line with recommendations from Lakens, 2022). However, to gain an understanding of the equivalence and minimum effect tests’ power, we conducted sensitivity analyses assuming a range of plausible true effect sizes, which we describe on p. 32:

“We have planned the design to be able to detect the SESOI of d = 0.20 with 90% power. We conducted additional sensitivity analyses for the power of the equivalence and minimum effect tests. The power of both of these tests depends on the true effect size, and how close it is to the SESOI: If the true effect size happened to be identical to the SESOI, neither the null hypothesis of the equivalence test (i.e., an effect as large or larger than d = 0.20) nor that of the minimum effect test (i.e., an effect below d = 0.20) could be correctly rejected: every significant result would be a type I error. The closer the true effect is to d = 0.20, the more participants are needed for a high-powered equivalence and minimum effect test.

Assuming a true effect of zero, the one-sided equivalence test at the final stage of the sequential design would have 99% power (with n = 1092 and alpha = 5%). Assuming a true effect of d = 0.1, the equivalence test would have 75% power. Assuming a true effect of d = 0.35, the minimum effect test at the final stage would have 97% power (with n = 1092 and alpha = 5%). Assuming a true effect size of d = 0.30, the minimum effect test would have 76% power.”

Reviewer 2.6

Relatedly, do the planned minimum effect tests have any implications for the alpha spending function in the sequential analysis? (I’m not very familiar with sequential analyses myself!)

Response:

We thought extensively about this point and also asked Daniël Lakens what he thought about our combination of NHST, equivalence test, minimum effect test and sequential analysis. As we understand it (and Daniël Lakens confirmed), we use the alpha levels calculated by the alpha spending functions for both the equivalence and minimum effect tests at every stage of the sequential design. The conventional NHST will only be conducted only once, namely if at the final stage of the sequential design, both equivalence and minimum effect test are still inconclusive. In that case, we will also use the alpha level for the last stage from the Pocock like spending function for this test. Otherwise, we see no implications of the minimum effect test for the alpha spending functions.

Reviewer 2.7

I think the repeated measures design is appropriate for increasing power, but one risk it presents is that it could make it a bit easier for participants to guess the hypothesis. I don't think this is a crucial or unusual feature of the study, but perhaps this could be flagged as a limitation in the eventual discussion.

Response:

Thank you for this suggestion. We will describe this as a potential limitation in the discussion. We also try to reduce this risk by increasing the time between T1 and T2 (see also response to Reviewer 1.9).

Reviewer 2.8

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

There is a great deal of well thought-through methodological detail in the manuscript already, and this criterion is close to being met. However, it could be useful to resolve the following potential ambiguities:

I take it that all the materials presented to participants will be in German?

Although I appreciate that the conspiracy belief items for the RR are shared on the OSF project, given that these constitute the key DV it'd be useful to list all three in the manuscript itself rather than just one example (bottom of p. 19).

P. 20, please ensure that all measures that will be administered are listed ("such as" implies that there are other unstated measures).

Response:

Yes, all materials of the Registered Report and the pilot studies will be/were in German. We now mention this on p. 20 for the Registered Report, and on p. 10 for the pilot studies. Please also see our English translations of the study prototypes (as described in our response to the Editor):

T1: https://www.soscisurvey.de/regreport/?act=2ttwhD6lxBaQk70DtFA1JTr2

T2: https://www.soscisurvey.de/regreport/?act=MYwjyVcPtt8tvaPq6fmkbuhq

We now describe all DV items in the text on p. 21., and list all exploratory measures that will be administered on p. 22.

Reviewer 2.9

What payment per participant will be provided on Prolific? Will they receive separate payments for the baseline and main study surveys?

Will participants be paid if they meet exclusion criteria? (I'd suggest they should be, since the criteria Prolific have for rejecting submissions are very narrow, and don't really cover any of the exclusion criteria listed in this study).

Response:

Since we now use the SoSci Survey panel, participants will not receive payment. They will be able to participate in a prize draw of 5 vouchers worth 100 €.

Participants that meet the exclusion criteria (depression and suicidality screening, not participated seriously) will also be allowed to participate in the raffle.

Reviewer 2.10

P. 21, brooding manipulation – if participants complete all 7 questions twice, will they then be presented a third time? And what happens after 5 minutes – does the page automatically progress even if the participant is still typing?

Response:

We have planned this in the following way: everyone in the brooding and reflection condition will go through the first cycle of questions + one cycle of repetitions. Then, the repetition questions will begin one by one again. At every page, it will be assessed whether, by now, five minutes have passed. As soon as five minutes have passed, the "continue" button will bring participants to the DV (instead of to the next question).

Reviewer 2.11

P. 23, re. the question "What were your thoughts in the 5 minutes before we asked you the questions about the conspiracies?" Why not just put this question before the conspiracy ones, and say "..in the last 5 minutes"? Perhaps I'm missing something here.

Response

We deliberately placed the manipulation checks after the dependent variable. We think that answering the MCs first might influence the DV: the questions about how they thought about the worry topic (whether they brooded/reflected etc.) might introduce a reflective meta-mindset that could disrupt participants' brooding state and interfere with the manipulation. We also want to keep measurement of the MC independent from the DV, so participants should not include their thinking about the DV into their answer to the MC items (which is why we use the somewhat complicated wording of "before we asked you about the potential influence of secret, powerful groups"). See also our response to Reviewer 1.11.

Reviewer 2.12

P. 24 – "This means that we will consider our effect practically meaningful if the confidence interval of the effect size estimate is beyond d = 0.20, and practically negligible if it is below d = 0.20." Does this mean that you will consider the effect practically meaningful if the lower limit of the 90% confidence interval for d is greater than +0.20? And practically negligible if the lower and upper limits of the 90% CI for d both fall within the range -0.2 to 0.2? In general, I'd encourage the authors to be a bit more explicit and rigid in how they state inferential criteria in the main text, though Table 5 is very helpful in this regard.

Response:

Thank you for this helpful comment. We have attempted to be more consistent in how we describe the inferential criteria throughout the manuscript. We will consider the effect practically meaningful if the lower limit of the 90% CI falls ***above*** *d = 0.20, and practically negligible if the upper limit of the 90% CI falls* ***below*** *d = 0.20 (since we do one-sided tests).*

Reviewer 2.13

Could you please explain how you will apply the sequential sampling plan on Prolific? E.g., will you have just one study running, but pause it at particular points? Or have 3 separate study advertisements in a row, but exclude participants from part 2 if they participated in part 1, etc.? If the latter, be a bit wary with this – Prolific does have a prescreening function for excluding participants from prior studies you've run, but in my experience it doesn't seem to work very predictably (e.g., when I apply it, the available sampling frame doesn't seem to change). You might choose to use a "custom blocklist" instead or in addition.

Relatedly, how in specific will you run the baseline step? Does a 1/3 sequential block involve baseline plus main study survey, or will you first run a very large sample of baselines and then only the second part in blocks?

Response:

We will implement the design with the SoSci Survey Panel as follows:

In the first step, approximately 1,000 participants will be recruited for T1. These participants will be invited to T2 5-10 days later. Those who did not pass the depression or suicidality screening, or did not complete T1 until the end, will be filtered out in the beginning of T2. We hope that, from this first round of invitations, about 820 participants (i.e., about 50% of the full sample) will complete T2 and pass the exclusion criteria. If this is not the case, more participants will be recruited, until about 820 can be included in T2. If no conclusive result is obtained after this first stage, another batch of 800 participants will be recruited for T1, and will be invited to T2 5-10 days later. If this is not sufficient to achieve the full sample (N = 1638), more participants will be added successively until the full sample size is achieved.

Reviewer 2.14

How will missing data be dealt with? Please consider both whether there might be a need to exclude participants based on quantity of missing data, and also how you might handle participants who are included in the final sample but have missing data on some items.

Response:

All main items will be mandatory (participants will not be able to continue the survey unless they give an answer to the manipulation check and DV items, in total 12 items), so there will not be participants with missing data on these items.

We plan to use only those cases with complete data on the dependent variable and manipulation checks. So, everyone who stops the survey before completing the MC items will not be included. Participants will also have the option to cancel the survey and demand that their data are deleted. Those who chose this option will also be excluded (but, if they wish, will be manually entered into the raffle).

Reviewer 2.15

Could the authors confirm that all the t-tests mentioned in the analysis plan are Welch's t-tests?

Response:

Yes, all t-test will be Welch's t-test, as we now mention on p. 30.

Reviewer 2.16

Could you please add a bit more clarity about the sequential analyses? I.e., what will the alpha level be in the first step? The second? The third? (Most readers won't have used sequential analyses before, so try to step us through the details slowly!)

Response:

We have tried to describe the sequential analysis more clearly on p. X. Since we now use the SoSci Panel instead of Prolific, we now use a sequential design with two looks instead of three (see our response to the Editor's comments):

"Using the rpact package (Wassmer & Pahlke, 2022) we have designed a sequential study with 90% power for d = 0.20 in a one-sided test, an alpha level of 5%, and two equally spaced looks (the first look after approximately 50% of data have been collected). The Type I error rate is kept at 5% across both looks using a Pocock-like alpha spending function, and the Type II error rate is kept at 10% using a Pocock-like beta spending function.

An a priori power analysis shows that at most 546 participants per condition are needed (total N = 1,638). The first look will be after approximately 820 participants have been collected.

Using the Pocock like alpha spending function, we can calculate the alpha levels at each look that will lead to a rejection of the respective null hypotheses of equivalence, minimum effect and conventional t-test. At the first look (50% of data), the alpha level is.031. At the last look (100% of data), the alpha level is 0.30.

When there are deviations from the pre-planned number or timing of looks, the alpha spending function allows to recalculate the alpha levels based on the exact amount of information that has been observed. So, it is not strictly necessary to analyze the data exactly after 50% have been collected (Lakens et al., 2021)."

Reviewer 2.17

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).

I think this criterion is largely met. The authors have been especially thorough in planning their manipulations check. A few minor comments:

P. 25, "So, we would consider an effect greater than d = 0.30 meaningful". Does this mean that the lower limit of the 90% CI of the effect in the manipulation check has to be above 0.30 for you to consider the analysis to be a valid test of the hypothesis? If this criterion isn't met, what will you do? Not report the main hypothesis tests at all? Report them but interpret them as being inadequate tests of the hypothesis? Something else? I do think there is a non-trivial chance that this manipulation check will "fail", so it's worth really thinking through this scenario. Relatedly, consider that for this manipulation check to have a good chance of "passing", the true effect size needs to be quite a bit larger than 0.30…

Response:

This is an important point. Yes, we suggest that the lower limit of the 90% CI of the effect in the manipulation check needs to be above 0.30 for us to be able to validly test the main hypothesis. Based on the results of Pilot Study 3 (which yielded fairly large effect sizes for the manipulation checks), we believe this is a suitable criterion. If this manipulation check fails, we will nevertheless describe the test of the main hypothesis and the overall pattern of results, but not consider it as evidence for or against the hypothesis and not draw any confirmatory conclusions (we now mention this on p. 29).

Reviewer 2.18

P. 26, "That is, we will consider the brooding manipulation effective if (1) the brooding condition scores meaningfully higher (at least d = 0.30) on the brooding MC than the control group AND (2) the control group scores meaningfully higher on the 'thinking as usual' MC than the brooding condition." I take it these criteria relate to the lower limits of a 90% confidence interval for d, not the sample estimates of d?

Response:

Yes, we meant the lower limits of the 90% CI for d, and clarified this in the manuscript.

Reviewer 2.19

I appreciate the attention paid to exclusion criteria, but I'm a little sceptical of the exclusion criteria relating to "nonsense" responses. Consider that participants can *only* be excluded based on these criteria if they're in the brooding or reflection conditions (people in the control condition won't be answering these questions). This means this exclusion criterion is confounded with condition, and applying it could bias the resulting estimates of causal effects (e.g., by removing the least attentive participants in the sample from two conditions, but not the other). I'd tend to think that it might be safer to leave out this exclusion criterion in your main analyses, but consider applying it for supplementary robustness analyses.

Response:

This is a very important point. We agree that this could create confounding and left it out for the main analysis.

We also checked whether dropping this criterion (which we had already applied in a similar way in Pilot Studies 2a and 2b) would change the pattern of results in the pilot studies. This was not the case.

Reviewer 2.20

Comments on pilot studies

A relatively unusual feature of this Stage 1 RR is the quantity of existing empirical work presented (4 pilot studies). These studies were described very briefly in the main text, and then in more detail in the supplementary file. This was very impressive; I really appreciate the work the authros were willing to do by way of preparation for the RR.

It nevertheless took me a while to get my head around what took place in all these studies and how they relate to the RR, which was a bit of a challenge given the scheduled review format. One thing I puzzled over was the salience of the fact that three of these studies were preregistered, but not in a great deal of detail (the aspredicted template isn't great for producing detailed registrations), and the reported analyses deviate from the preregistrations somewhat, without particularly detailed rationales. Usually, these would be things I'd grill authors about in a peer review! Yet what's different in this case is that the authors seem *not* to be using the preregistrations to claim any special confirmatory status for the studies as tests of hypotheses, but are instead presenting them as tentative pilot studies that are used to refine the methods and main hypothesis. In that context, perhaps the deviations from the preregistrations are less important? I'd be interested in what the authors think about this topic, and also Chris's take! Perhaps it's something that could be touched on, albeit briefly, in the main text (currently I don't think the preregistrations are mentioned there).

Response:

We thank the reviewer for going through all the material so carefully! The purpose of the preregistrations for the pilot studies was mainly twofold: First, we used the aspredicted forms as a useful tool to create a somewhat standardized snapshot of the main study idea to facilitate communication and planning. Second, we see these preregistrations as a very reduced form of "lab notebook" that documents the development of this line of research. Specifically, we developed the study concept while collecting data. Even though we preregistered confirmatory tests very early on, we did so with somewhat limited confidence in both the justifications of the predicted outcome and the adequacy of the study design (because we could not rely on previous research, be it by others or ourselves). For Pilot Study 3, specifically, we decided against a preregistration because we realized that our earlier preregistrations might have been somewhat premature.

After each study, we reevaluated the theoretical rationale and further developed the experimental design. We now mention the preregistrations in the main manuscript and describe any deviations transparently in the Supplement (where we also report results of the preregistered analyses). Our goal is to be as transparent as possible. Please let us know if you have any further recommendations.

Reviewer 2.21

Beyond that, I have some more minor comments on the pilot studies, relating primarily to content in the supplement:

I assume all pilots were conducted in German?

In the supplement, please state all measures administered in the pilot for the sake of transparency, even if not all were used in the analyses presented here.

Response:

Yes, we all pilots were in German. We now list all measures administered in the pilot in the Supplement.

Reviewer 2.22

For replicability, please indicate which survey company or companies recruited participants.

For replicability, please indicate how participants were compensated.

Response:

We included this information in the Methods part of each pilot study (Supplement).

Reviewer 2.23

In pilot study 2a, the significance test of the manipulation (i.e., effect on perseverative thinking) has p = .048 given a one-sided test, but the preregistration doesn't say the test would be one-sided (and it wouldn't be significant if 2-sided). Considering also that p values near 0.05 are only weak evidence against the null anyway (Benjamin et al., 2018) it might be worth noting that result provides only quite tentative evidence of an effect.

Response:

We have added this information in a footnote in the main text on p. 12: "Note that the p-value for this one-sided test was close to .05 (specifically, .048), and can thus only provide tentative evidence of a successful manipulation (Benjamin et al. (2018)."

We also mention this in the Supplement where we describe pilot study 2a in greater detail: "However, the p-value for this test is very close to 0.05. Therefore, it is important to note that this result can only provide tentative evidence for a successful manipulation (Benjamin et al., 2018)."

Reviewer 2.24

The prereg for Study 2a also says "We will check assumptions of normality and homogeneity of variance. If not violated, we will conduct student's t-tests to compare means in conspiracy beliefs between rumination and control conditions for each scenario separately. If assumptions are violated, we will conduct non-parametric tests of mean differences." That doesn't link particularly closely with the Welch's test reported, which is a reasonable choice but neither a Student's t-test nor (strictly speaking) non-parametric. In general, I think it'd be useful in the supplement to signal and explain deviations from the preregistrations in a bit more detail, even if these deviations may be less crucial in the context of these being pilot studies than if they were being presented as confirmatory.

Response:

We realized that the Welch's t-tests made more sense than either non-parametric or Student's t-test, based on recommendations by Delacre et al. (2017). We describe this (and other deviations from the preregistrations) in the Supplement. Each preregistered pilot study now has a section "Deviations from preregistration".

Reviewer 2.25

Pilot study 2b, "The main goal of this study was to replicate the finding from pilot experiment 1". Does the latter refer to pilot Study 2a?

Response:

We thank the reviewer for spotting this typo (now fixed).

Reviewer 2.26

Pilot study 2b, "Given that the manipulation failed to exert a significant influence on state rumination, we neither expect an effect on conspiracy beliefs" – I understand what the authors mean, but it's a bit of an odd phrasing given that the preregistered hypothesis said otherwise.

Response:

We agree that this phrasing was confusing and changed it. It now reads: As such, we cannot consider the manipulation successful. The rumination condition also did not report significantly higher conspiracy beliefs than the control condition (...)."

Reviewer 2.27

Am I right in inferring that pilot study 3 was not preregistered? (This seems fine, but perhaps mention this, since the others were)

Response:

It is correct that this study was not preregistered (see also Response 2.20). We now mention this in the Supplement.

Reviewer 2.28

I fully appreciate that the authors will be aiming to keep the descriptions of the pilot studies in the main text brief so as to meet word limits at PCI-friendly journals, but perhaps a tiny bit more info about fundamental issues like country of sample, language and recruitment method might be helpful to readers who look only at the main text.

Response:

This is a helpful suggestion. We have included information about the country of sample, language and recruitment for the pilot studies in the main text on p. 10, 11 and 14.

Reviewer 2.29

Comments on introduction

I thought the introduction was excellent, but just have a few very minor suggestions for clarifications.

Defining conspiracy beliefs subsection: "A conspiracy is a secret plot by a powerful group that aims to achieve a common goal. Importantly, the conspirators pursue their goal without any regard for other people or consequences for society as a whole." This latter part of the definition is quite strong and a bit unusual relative to conventional definitions of "conspiracy", which tend only to require that the conspirators are plotting to harmful or malevolent ends (one can have some regard for others while still doing something harmful). C.f. a dictionary definition. The authors aren't "wrong" for using this definition (definitions are always socially constructed) but I'd be interested to know why they prefer it!

Response:

We deliberately chose this latter part of the definition in favor of the conventional phrasing of "malevolent intentions". We believe that the conspirators need not commit their actions explicitly for the sake of harming others, that is, they need not have the intention to cause harm. It is sufficient if they do not care if their actions harm others, and act regardless of whether harm is inflicted. We believe that in most cases, people do not act simply for the sake of being evil, but because they have something to gain from their action and accept the negative consequences that this has for others.

We adapted the definition to reflect this more clearly:

"A conspiracy is a secret plot by a powerful group that aims to achieve a common goal. Importantly, the conspirators pursue this goal regardless of the consequences for others: Malicious intentions are not required, but the goal is pursued even if others are harmed."

Reviewer 2.30

"Conspiracy beliefs have harmful consequences for individuals and societies" – I'd suggest softening this a bit to be more consistent with the rest of the paragraph, where you acknowledge that it "lies in the public interest to uncover true conspiracies".

Response:

*We agree with this point, and changed it to "It lies in the public interest to disprove false, and uncover true conspiracies, particularly because conspiracy beliefs **can** have harmful consequences...".*

Reviewer 2.31

"Studies have shown that rumination increases depressed mood…" – Were these experimental studies (thus justifying this causal inference)? In general, you could perhaps give just a little more information about the methods of key studies you're reviewing, so that the reader has some basic impression of the strength of the evidence in these prior studies.

Response:

We have taken a closer look at the literature on rumination and negative affect. Indeed, many of the studies are experimental, but they have some limitations which we now describe more transparently. In this process, we also came across some relevant literature we had missed before, which is now included on p. 6-7.:

"Crucially, it is well-established that rumination in response to distress increases negative affect. Rumination has been described as an "emotional magnifier" that amplifies existing negative affective states (Watkins & Roberts, 2020, p.2). A number of experiments have shown that ruminating about distressing events prolongs negative mood. These studies have typically used a repeated measures design in which a rumination condition was compared to a distraction condition, and negative affect was measured before and after the manipulation. In a comprehensive review of research on the link between rumination and negative affect, Kirkegaard Thomsen (2006) concluded that 15 out of 20 studies that used such a design found the predicted group difference between rumination and distraction, two reported a trend in the expected direction, and three reported null results (which may, in part, be attributable to

failed manipulations). However, these studies did not examine whether effects resulted from an increase in negative affect due rumination, or a decrease in negative affect due to distraction (Kirkegaard Thomsen, 2006). As such, one can conclude that rumination increases negative affect compared to distraction, while its effects alone are less well studied experimentally.

Beyond these experimental studies, a number of longitudinal investigations provide evidence for a link between rumination and negative affect: the tendency to ruminate has consistently been found to predict longer and more severe periods of depression at a later time (Nolen-Hoeksema et al., 1997; Nolen-Hoeksema et al., 1994). Similarly, a recent experience-sampling study found evidence for a reciprocal relation between rumination and negative affect: within-person increases in rumination predicted subsequent within-person increases in negative affect, and vice versa (Blanke et al., 2022). Converging findings have been obtained by researchers using similar designs (Brans et al., 2013; Lennarz et al., 2019; Moberly & Watkins, 2008; Pavani et al., 2017)".

We also came across some interesting recent research on rumination and negatively biased thinking, which we included on p. 7-8:

"A more recent study using a thinking-aloud paradigm further found that participants with higher trait rumination scores (specifically, trait brooding scores) demonstrated longer periods of negative thoughts in a resting state, and their negative thoughts were linked to a stronger narrowing in conceptual scope over time, as indicated by higher semantic similarity (Raffaelli et al., 2021). This converges with Andrews-Hanna et al. (2022)'s finding that, during a free association task, trait ruminators are more strongly attracted to negative conceptual spaces and more likely to remain there longer."

Reviewer 2.32

"Considering that conspiracy beliefs tend to emerge when people experience negative affect…" The phrasing here could be read as implying that there is *empirical evidence* that conspiracy beliefs emerge when people experience negative affect, whereas I think you're just intending to say that this is what some *theories* assert. As you'll know, the empirical evidence that negative affect causes increased belief in CTs is somewhat tentative.

Response:

This is a good point. We rephrased the sentence: "Given that theories on the formation of conspiracy beliefs state that they are more likely to emerge when people experience negative affect (Douglas et al., 2017; van Prooijen, 2020; van Prooijen & Douglas, 2018).

Reviewer 2.33

Lastly, my apologies for any typos or lack of clarity in my review; given the scheduled review format I thought it was more important to do this quickly than to carefully draft a very elegant review document! Please likewise forgive me if any of the points or questions I've raised above are already covered in places I've missed.

Response:

We would like to thank Matt Williams very much for their detailed comments and for conducting the review so quickly and thoroughly. Their comments were a great help to us.

References

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Lakens, D. (2022). *Improving your statistical inferences*. https://lakens.github.io/statistical_inferences/index.html

Mayo, D. (2022). *Do "underpowered" tests "exaggerate" population effects?* https://errorstatistics.com/2022/05/02/do-underpowered-tests-exaggerate-population-effects/

Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Cognitive Psychology* (pp. 4–31). Psychology Press.