Dear Dr. Zhang,

Thanks for inviting us to revise our submission. We have marked all our edits in this submission using TRACK CHANGES.

Sincerely,
Authors.

Table of Contents

Editor

Dear Authors,
Thank you for your efforts in this revision and for your patience during the review process. I have now received all three reviews. Based on these reviews and my own reading of your manuscript, I would like to invite you to revise the proposal.

As attached below, the reviewers appreciate your engagement with their comments and the revisions made, but they remain somewhat critical of certain aspects of the proposal. I also agree with the reviewers that there seem to be some misunderstandings between the authors and reviewers, as certain responses do not fully address the points raised (e.g., the randomization of condition assignment versus the randomized order of scenarios). I find the reviewers' insights helpful and hope you will engage with their comments carefully and thoroughly.

In addition to the reviewers' comments, I would like to raise a few points myself:
Before commenting on the specifics of your revision and response, I have one general comment on the submission as a programmatic RR. After deliberation with the Managing Board, I reached the conclusion that the proposal does not meet the criteria laid out in the author instructions for programmatic RR (https://rr.peercommunityin.org/help/guide_for_authors#h_52492857233251613309610581). The dependency of results on S1 and 2 leaves too much freedom in the design and analysis for Study 3. And it is difficult for me to see that Study 3 could be published as a stage 2 report without further reviews on its methodology. I therefore, strongly suggest you to remove study 3 from the current stage 1 report and later submit study 3's protocol as a separate submission. I would be happy to handle the next submission to increase the efficiency of the review process.

Duly noted, we agree with the decision that the present proposal does not qualify for programmatic RR status. We have removed references to Replication 3 and will submit that proposal later.

Regarding one reviewer's suggestion on the design, I believe the difference between the two approaches is relatively minor (e.g., sample size in different conditions). Essentially, you have three conditions in two experiments: 1) high specificity only, 2) low specificity only, and 3) both types. The reviewer's suggestion seems more focused on conceptualization and analysis rather than experimental mechanics. Therefore, I am not entirely convinced by your reply. While the original analytic plan of examining the data separately is reasonable, it may still be valuable to compare the data from the two experiments as the reviewer suggested. I encourage you to give this some more consideration.

Please see response to the comment below which includes the metamodel recommended here.

Regarding the revision to Hypothesis 1, I believe the reviewers are not questioning your judgment on cognitive interviews but rather highlighting that the revised hypothesis is not well-supported by the literature. Thus, it may not be best practice to treat it as such. In my reading, including this hypothesis in the introduction may not be necessary. If subsequent analyses reveal no preference for pragmatic correspondence, you can still discuss its implications for current theory in the Stage 2 report and explain how this result aligns or diverges from other research or theoretical frameworks. Additionally, I did not see the expected results if the revision hypothesis is correct in the study design table. I assume there should be a non-significant result regarding question type and a positive/negative intercept suggesting a bias in all responses?

Many thanks for flagging this issue. First, we present the Revision Hypothesis as a potential outcome (given what we found in previous research [Neequaye & Lorson, 2023]), not as an "established" result. For that reason, we believe an explicit prediction and a corresponding tests are warranted to ensure robustness and continuity in the current research program. Second, we have now reframed Revision Hypothesis 1 into Revision Hypothesis 1a and 1b to make the testing and expected outcomes clear (please see Present Research and the Study Design Template). The Revision Hypothesis is not merely the inverse of the Core Hypotheses. Revision Hypothesis 1a tests the prediction that High-versus low-specificity questions do not elicit more designations of information items that align with pragmatic correspondence. And the corresponding result should be that the Question Type parameter's HDI is predicted to fall within the null region, such that we can conclude the data are consistent with 'no effect' of question-type. Revision Hypothesis 1b (which is the meta-model you recommended) tests the prediction that High- versus low-specificity questions manipulated as a between-subjects versus within-subjects factor do not elicit more designations of information items that align with pragmatic correspondence. And the corresponding result should be that the interaction parameter of question x design type HDI will fall within the null region, such that we can conclude the data are consistent with 'no effect' of question type x design type.

Regarding the analytic approach, it is important to recognize that not all colleagues are familiar with every analytic method. While the detailed reporting of modeling decisions is commendable, it may overwhelm some readers, especially those unfamiliar with Bayesian methods. I suggest that in the later report, you also present descriptive data to make the main findings more accessible. I appreciate your engagement with the reviewer's comments by adding the rationale for the analytic approach, but in my opinion, the rationale is somewhat general and does not address the real issue: the accessibilty of results. For instance, the mention of interaction and confound control seems less relevant based on my reading

of the analysis plan. Additionally, the advantage of Bayesian statistics in small samples is less apparent here, given the relatively large sample size. As such, I believe this discussion may not be necessary.
I agree with the reviewers that the coding of participants' responses could use some further clarification. I also checked the codebook but there isn't much explanation other than some example entries so I could not make sense what would the coding process look like (my apology if I have misread the file).
Regarding the optional wager, could you please clarify whether the wager was hypothetical or involved real monetary stakes? If it was purely hypothetical, I have concerns about its validity as a measure, as participants may provide random answers without real stakes.

Thanks for raising these issues. We will clarify them in the chronological order.

1. **Accessibility of the analytic method:** We completely agree with the sentiment and will make the results and discussion sections accessible by presenting descriptive data as well as inferential statistics. **In that regard, should we delete the rationale or leave it for now**?
2. **Codebook:** On page 14, we describe how the coders will execute their codings. They will rate the extent to which high- or low-specificity would better elicit (or better fit) the responses participants provide. At the osf link, we provide a data dictionary with which anyone can reproduce the coding instrument and independently code the data—suppose an independent analyst wants to scrutinize the findings.
3. **Wager:** As noted on page 13, the wager is hypothetical and optional (participants could skip it). Because of the possibility that the validity of the wager could be limited, we implemented it as a secondary measure to the confidence rating (i.e., the wager is NOT the main dependent variable, confidence is the main one). The wager is an *additional* measure to explore another dimension of confidence. We are open to reporting that measure as exploratory in the results if you insist.

One reviewer suggested including exploratory analysis in the Stage 1 report. While this is common in some journals, I would recommend leaving any exploratory analysis out of the Stage 1 report to maintain a clear distinction between confirmatory and exploratory analyses.

We completely agree and will report any explorations for consideration at Stage 2 (if the manuscript makes it there).

A final minor point. While I understand that it was likely not your intention, some expressions may come across as confrontational in the response letter. For example, in response to Reviewer 1's comment on analysis, you stated 'Respectfully, the Analysis Strategy section is not for the lay reader but for an expert to assess.'. In addition, I imagine that there could be a better way to convey your point than the inclusion of the screenshot of dictionary. In written communication, misunderstandings can arise even with the best intentions. I kindly request that you make minor adjustments to the tone in future communications to avoid potential misinterpretations.

This point is well-noted. Apologies if some responses sounded confrontational. Our intention was only to communicate as explicitly as possible, leaving no stone unturned. We will ensure check our tone in future responses.

Best regards,
Yikang

Review by anonymous reviewer 1, 19 Sep 2024 10:29

Thank your for the opportunity to review this report again. I appreciate the additions the authors have made to the report - it is much clearer now. However, I believe that my main points have gone unaddressed. While I am not going to reiterate them here, let me point out the following:

Many thanks for reviewing our revision. We appreciate your time and comments (especially concerning the need to be explicit about testing the revision hypothesis).

It is indeed true that a study that sets two theoretical models from which two competing hypotheses have been derived and where the data will then support one of these hypotheses is optimal for scientific progress. I do not believe that this is the case here. The authors merely state that if the main hypotheses are not supported, then the theory needs to be revised. This is quite standard and for this reason the revision hypotheses do not seem to be necessary.

We have now reframed Revision Hypothesis 1 into Revision Hypothesis 1a and 1b to make the testing and expected outcomes clear (please see Present Research and the Study Design Template). The Revision Hypothesis is not merely the inverse of the Core Hypotheses. Revision Hypothesis 1a tests the prediction that High-versus low-specificity questions do not elicit more designations of information items that align with pragmatic correspondence. And the corresponding result should be that the Question Type parameter's HDI is predicted to fall within the null region, such that we can conclude the data are consistent with 'no effect' of question-type. Revision Hypothesis 1b (which is the meta-model you recommended) tests the prediction that High- versus low-specificity questions manipulated as a between-subjects versus within-subjects factor do not elicit more designations of information items that align with pragmatic correspondence. And the corresponding result should be that the interaction parameter of question x design type HDI will fall within the null region, such that we can conclude the data are consistent with 'no effect' of question type x design type.

I also continue to believe that the analytical approach is overly complex for the analytically simple questions that are being asked.

This comment is a fair one, and we accept the disagreement. As suggested by the recommender and noted previously, we will make the results and discussion sections accessible by presenting descriptive data and inferential statistics.

Finally, the authors state: "Before an interviewer poses any question, it is reasonable for interviewees to assume that the elicitation of complete details is the de facto purpose of an interview. All things being equal, any investigator would want the complete details an interviewee holds, given that complete information would be more beneficial to any investigation than partial details." I do not believe these can be assumed without accompanying evidence. I would not agree with them in any case and they seem to go against the pragmatic principle of optimal relevance in communication.

This comment is also a fair one, and we agree that such a claim requires an explicit test. The revision hypothesis, whose test we have now made explicit, will examine our claim that "it is reasonable for interviewees to assume that the elicitation of complete details is the de facto purpose of an interview." We have now reframed the Revision Hypothesis to make it clear that the hypothesis tests the claim highlighted here.

Review by Feni Kontogianni, 23 Sep 2024 16:20

The authors have responded to most of my comments/questions. Below I outline a few remaining issues that are straightforward to address. In fact, for all except the first comment below, the authors responded in full in the response letter and I simply ask that they add these statements in the manuscript too.

Thank you for your time and feedback. We appreciate your thoughtful comments (especially regarding the potential to mistake cooperative interviewees for reluctant ones when their memories fail).

I made the following comment before: 'in practice cooperative interviewees may understand that the interviewer wants highly specific/complete details but may be unable to report these from memory. Some consideration to the implications of the current assumption risking cooperative interviewees being perceived as uncooperative is warranted'. The authors replied stating that the focus here is about pragmtic considerations and not research on memory or disclosure. Fair, and I understood the authors' thinking. My comment is about considering the practical implications of their reserach findings - and any unintended consequences. As the authors state: ''By asking high- or low-specificity questions, interviewers can influence the extent to which interviewees might discern what they want to know" and depending on the findings it is plausible that interviewers can assume that their objective is clear and without awareness of memory issues wrongly infer the interviewee's response means they are being uncooperative when they are unable to remember - and saying so would not necessarily make a difference. If this seems too obvious, I would say that it is still worth addressing especially since it can very easily be addressed by adding a few sentences in the introduction where the authors can both emphasise that disclosure based on memory is not the current focus and clarify that factors based on memory can prevent cooperative individuals from providing specific information even if they recognise the objective. The point could also be considered in the discussion depending on the findings as a caution, but it is up to the authors at that point.

First of all, accept our sincerest apologies for the oversight. We indeed missed the point you raised, which is crucial and highlights the need for this line of research. We have now expanded the discussion on page 6 (see TRACK CHANGES), commenting on the potential for interviewees to be mistaken as uncooperative if they acknowledge an interviewer's objective but cannot provide the information the interviewer wants.

- The authors addressed another of my comments by stating that in the current reserach they 'simply claim that one must understand a question's purpose before disposition or other contextual factors most proximate to disclosure can take effect. Identifying the question's purpose has an indirect influence because it tells the cooperator or resistor how best to achieve their goals. Deciding what to do, in this case disclosure or nondisclosure, is when disposition kicks in'. Given how clear this statement is including it in the manuscript would be ideal to avoid confusion.

Noted with thanks. We have now included the highlighted explanation on page 5.

- Regarding comment 5, I appreciate that the authors have included detailed information in the Study design template. My mistake, I should have clarified that I was referring to the first mention of this statement on p. 9-10, and my comment referred to clarity in text so that it is easier for the reader to follow. The authors have now added a clarification on this point although it might be worth using very direct language i.e. participants will be more likely to identify that the interviewer aims to focus on highly specific details.

Thanks for the clarification. We have now framed the hypothesis more directly, noting that pragmatic correspondence refers to specific details.

- ICC value and coding strategy: yes, it was clear that the authors only referred to the possibility of the ICC value being below 0.6. To clarify my point was that if the ICC is below this value then preferably the authors would report on this in the final manuscript to some (appropriate) extent as it would be a limitation. The reason I raised this is because they state that they would 'discuss the causes of disagreements and resolve them independently' before recoding the data, but ideally some more detail about the issue would be reported. Hopefully this is clear and ideally not even needed in the final manuscript.

Noted with thanks. We would report all aspects the coding process to ensure transparency.

- In the response letter, the authors state that they will explore all the data points that receive the designation 'cannot decide' for any obvious patterns and label them as exploratory. I would then suggest that they mention this already in the manuscript too, before it is accepted for submission.

Duly noted. We defer to the recommender on this issue about the inclusion of exploratory analysis at Stage 1. If the recommends agrees that we mention exploratory analysis at Stage. The we will do so.

Review by anonymous reviewer 2, 08 Oct 2024 21:08

I appreciate the authors attention to my comments. Below I have reproduced sections of my previous review where I felt the issue wasn't sufficiently addressed (the original text of my review is in regular text and my new text italics). I have deleted comments I felt have been addressed in this revision. At the bottom of all that I have added a few other comments, in italics.

Thank you for taking the time to review our revision to help us improve.

FROM BEFORE: I found the hypotheses to be appropriately precise, and sufficiently conceivable as to be worthy of investigation. That said, I found the wording of "preference for pragmatic correspondence" to be confusing, as the outcome isn't the participants' preferences. A slight rewording would be helpful. NEW: It seems my comment above may have been misunderstood. What I was trying to say is this hypothesis makes it sound you are asking people for their preferences. That is not the case. So, I would reword this to better reffect what you are measuring. In this study, the primary outcomes is whether their response is scored as pragmatically corresponding more to either the high- or low- specificity question. So, the language should align with that.

Thanks for clarifying and apologies for the misunderstanding. We have now reworded Hypothesis 1 in the main text and the study design table to reflect that high- versus low-specificity questions should significantly influence the perceived specificity of interviewees' responses.

FROM BEFORE: On page 6 I was a bit thrown by the paragraph starting "Before an interviewer proposes any question, it is reasonable for the interviewees to assume that the elicitation of complete details is the de facto purpose of the interview." This seems to argue against their first core hypothesis. If this is being presented as an alternate theory that didn't fully come across.

NEW: I appreciate the clarification that this is indeed relevant to the revision hypothesis. But that still doesn't come through in the proposal.

Thanks for raising this point. We are glad that our clarification is now evident. We mention the competing hypothesis in the abstract. Additionally, we are also now explicit about the pattern of results that would support the revision hypothesis. Revision Hypothesis 1a tests the prediction that High-versus low-specificity questions do not elicit more designations of information items that align with pragmatic correspondence. And the corresponding result should be that the Question Type parameter's HDI is predicted to fall within the null region, such that we can conclude the data are consistent with 'no effect' of question-type. Revision Hypothesis 1b (which is the meta-model you recommended) tests the prediction that High- versus low-specificity questions manipulated as a between-subjects versus within-subjects factor do not elicit more designations of information items that align with pragmatic correspondence. And the corresponding result should be that the interaction parameter of question x design type HDI will fall within the null region, such that we can conclude the data are consistent with 'no effect' of question type x design type.

FROM BEFORE: It's not clearly stated that random assignment will be used for the between subjects manipulations (I rather assumed it would be, but it should be stated in the text/procedure).
NEW: I appreciate the authors' attempt to address this, but what I was looking for is a statement that the participants will be randomly assigned to their between subjects condition. That is still not clearly present. The reference to the "5 randomized scenarios" I took to mean the order would be randomized.

Noted with thanks, we have now added an explicit sentence that participants will be randomly assigned to the high- or low-specificity condition.

FROM BEFORE: It wasn't clear what the "decision making" manipulation check was (in Exclusion Criteria section) but I assume it refers to that "instructional" manipulation check (in Appendix B).
NEW: I appreciate the response, but I think clarity would be greatly improved by simply referring to it as the instruction manipulation check, rather than "decision making" and/or an acronym. There is no need for an acronym.

Noted with thanks, we have now deleted "decision making" to improve clarity.

FROM BEFORE: The textbox prompt "The police-contact wants to know if…" seemed odd to me. Specifically, the word "if." That seems like it would generate answer like "…if I know what brand of drugs the gang is selling" as opposed to answer like "…the gang is selling off brand oxy" (I assume the latter is the type of answer the authors' are seeking. I realize the authors piloted this, so I'm willing to defer to them on this point. It just seems very odd to me.
NEW: The response to this was very enlightening. I didn't come away from this proposal at all understanding that the desired responses were "They want to know if I know XYZ." I realized that they were not supposed to put what they would actually report to the interviewer, but I still thought actual content was what you wanted. E.g., "They want to know the gang hangs at exit 7F" (Note, I'm not saying that would be what the participate would TELL them; it's clear that's not the goal. But I thought that you wanted the participant to write out the content that they thought that interviewer wanted to hear.) Providing examples of potential responses in the method section would have been wildly helpful to make more concrete what types of responses you are trying to get at. The authors should add such examples.
Now that I understand the intent, two more concerns come to mind. First, it seems like the participants would just repeat back the question. E.g., "The interviewer wants to know if I know anything about the

gang's transportation in the park." If participants do that, I imagine the first core hypothesis would be confirmed. But I'm not sure it would be telling us much. Second, if participants are providing responses like "The interviewer wants to know if I know what brand of drugs the gang is selling." Then, this doesn't actually include any of the information (i.e., doesn't mention oxy, off-brand, green-star) to gauge whether the answer was "complete".

The new concerns are fair points. But in our pilot testing, we have found that participants do add specific details by which coders can gauge completeness (e.g., "if they sell off brand green star oxycodone"). The pilot data is publicly available to verify our claim.

FROM BEFORE The authors state that in their past study, and in R1 and R2, pragmatic correspondence was designed to be equivalent to complete details – so high specificity questions specifically request complete details. I thought I understood this, and it made sense to me. But once I got to the Appendices and saw the scenario information and the high and low specificity questions, I had some concerns. In fact, this is my biggest concern regarding the proposed studies.
NEW: I found myself in disagreement with the authors as relates to several points in their response. To ensure my understanding of their position is clear, I will start with that, in case I have misunderstood. Pragmatic correspondence (with the high specificity question) was designed to be equivalent to complete details. Using the first scenario, the authors argue that in order to provide the information corresponding with "Have you spotting the exact location at the park where KET22 deals drugs?" the participants need to mention all the information (edge of the park; discreet; exit 7F) in order for "Exit 7F" to be meaningful. I believe I understand what the authors are saying (though I disagree as explained below), though I'm not entirely sure when looking at each scenario what the 3 relevant details are. E.g., for the scenario about the contents of interactions, is the timing considered part of the complete info? Is the fact that it was an argument part of the complete info?
First, I just don't agree with the authors. The idea is that the high specificity question should (if their hypothesis is correct) create an expectation that a specific detail is sought, as per Figure 1. For this question the "specific detail" is the location. And the location is Exit 7F. The authors seem to be arguing, e.g., to give someone the location of your home, you'd have to mention not only the street and the house number, but also the contextual information that it's on a quiet tree lined street. Or, to use the example in Figure 1, to provide complete information it would be necessary to say not just 16:00 but also, in the evenings, after work. But that is inconsistent with Figure 1.
Second, in line with my comment above, if participants provide responses like "...if I know where in the park they deal" or "...if the gang deals in a specific spot in the park" – I don't see how that has anything to do with whether participants provide "complete" information, however that's defined.

This objection is challenging to wield as it contains a few moving parts, and we will defer to the recommender to decide as we disagree with the reviewer.

As noted, we have designed the scenarios (AND THE CORRESPONDING QUESTIONS) such that complete information will necessarily have to come with contextual details if the scenario calls for it. The example the reviewer raises about giving someone the location of your home **does not provide a corresponding question or context between the interlocutors**. As we theorize, it is the framing of a question AND the discussion context that determines the level of detail that would qualify as a pragmatically corresponding response (as we noted about the Exit 7F case in our previous response). So, the reviewer's objection here is unwieldy and leaves an overly wide goalpost by which to satisfy the reviewer. Depending on the question plus the context of the discussion (i.e., the level of detail requested), it may or may not be necessary to mention that one lives on a quiet street.

Note also that in Figure 1, the question asks for a SPECIFIC TIME, the answer of which is 18.00.

FROM BEFORE: I also have concerns with the fact that scenarios contain such limited information. To me this makes it reasonably easy for all participants to choose to provide all information. Unless I'm missing something, it seems like participants don't even need to be presented with any information. They could just be presented with questions and ask what information they think the interviewer would want them to find out. This is less leading as there are many potential options, not a couple details. Indeed, some of the low specificity questions are so vague that there are a huge number of details that an interviewee might suggest the interviewer was interested in, if they were not confined to 2 or 3 pieces of information. (e.g., "Have you discovered anything about the gang's narcotics sales lately?: This could be getting at whether sales good or bad; is one product selling better than another; sales are initiated via text messages; sales are primarily conducted by person X and person Y). As the proposed studies are more of an initial preliminary test, this is less of a concern than the previous point I made, but something to consider moving forward.
NEW: I'm not particularly persuaded by the authors response. I think my phrasing threw my point off a bit when I mentioned what the interviewer "would want to know." Perhaps a better example to make my point would be, what if there were a scenario where the participant was not able to learn anything that week, but the interviewer still asked a question. Much like someone could still make a judgment that the interviewer wanted to know about e.g., a bomb design, even if they didn't remember the answer, they could make a judgment that the interviewer wanted to know about a bomb design even if they didn't know anything about the bomb. Or, another example, what if there were a scenario where what they learned wasn't relevant to the question asked (they learned about where the bomb would be detonated, but not its design), the same basic logic would apply. They could make an assessment of the interviewer's goal despite not having relevant information. And then it wouldn't limit them to the contents of what they know (since they shouldn't be limiting their understanding of what the interviewer wants to know to the contents of what they know, as those things may or may not overlap). And, in line with previous comments, if responses like "the interviewer wants to know if I know about the bomb design" are the goal, then what does the interviewee's knowledge have to do with anything? All that said, since this is a preliminary test, I'll let it go. But I feel like what can be learned using the current design is fairly limited.

While we appreciate the point being raised here, we believe a misunderstanding still persists. Suppose an interviewer asks a question the interviewee had no knowledge about. In that case, the interviewee can correctly flag the objective and also flag that they (i.e., the interviewee) do not have the corresponding information. The question still leads the interviewee to focus on SOMETHING (however broad or specific)—and the interviewee does not have that information. **The interviewee's knowledge is relevant because that knowledge is how they (i.e., the interviewee) can make the judgment that they do not have the information the interviewer wants.** Granted, our study design is about when the interviewee does have the information the interviewer requests, which means interviewees can focus on something legitimate. So, we believe our design provides a decent test of the proposal and will provide more insight than a scenario where the interviewee does not have the information the interviewer requests. Because in a scenario where the interviewee does not have the relevant knowledge, the interviewee's can slice and dice the knowledge—that knowledge can only be used to make the judgment that they do not have the information the interviewer wants.

FROM BEFORE: The "disposition" manipulation was introduced in such a way on page 8 that I didn't realize it was a manipulation. I thought at first that participants could choose their disposition (also the term disposition was not introduced at that time, so it took me by surprise later).

NEW: The introduction disposition as a manipulation is much clearer, thank you. That said, it still comes out of nowhere. There should be some explanation of why this variable is there
/ what purpose is serves.

Noted, we mention disposition on page 5 and note why the variable should have no influence on the mental designation of information items.

FROM BEFORE: Coding: I was quite confused here. I think part of the is that perhaps where the authors said "choice' they really intended to say "response." If that is the case, I think I understand what the authors are proposing, but, it doesn't quite make sense to me. I think this is related to my major concern above. My understanding is that the authors are proposing that coder should code more "complete" answer on the "high specificity" side of the scale. But that assumes that the high specificity questions really were in fact seeking complete details. Which, as I note above, I don't think they (all) are.
NEW: In line with several points above, I don't really understand how this coding will work if the responses provided won't actually provide content and essentially repeat the question (perhaps some will and some won't?). Examples of a few potential response and how they might be coded would be exceedingly helpful. What would be an example of a response worthy of a score of +100 and -100? How would "...if I know the exact location where they deal" (which is basically repeating the question) score relative to "...if I know they deal at exit 7F" ? As the questions were piloted I assume the authors have some sense of what kind of answers they will get.
Also, will it be possible to catch if people are simply repeating the exact wording of the question?

Thanks for the comment, but we believe the comment overly anticipates that most participants will simply repeat the questions (and the objection is based on that assumption). As noted before, pilot testing indicated that prospective respondents do not simply repeat the questions. They do provide details. And coders can rate perceived specificity (i.e., what question best elicits a given response). We refrain from providing "potential responses" and provide the pilot results for examples of actual responses. If the editor insists, we will include examples of previous responses. In any case, the interrater reliability between the coders will show whether good consistency in perceived specificity can be achieved. A +100 response would be one that captures all the contents of a scenario (an issue the coders will determine and ICC will assess).

Other new comments:

Pg 5: It would be helpful to give a bit more elaboration on the statement "this honing process can indirectly affect whether an interviewee cooperates or resists" – just a brief example would help clarify the thinking here

Following a comment by another reviewer we have added more clarification to how we view the honing process and its consequences (see page 5)

Figure 1: I think it's a stretch to say this is showing the mechanism. It's more showing the basic procedure and the first core hypothesis. What the actual honing process/mechanism is, is not specified. The figure is really just showing the authors expect that if you ask for the specific time, people think you just want to know the specific time, and if you don't specify a specific time in the question, people don't make that assumption.

Thanks for the comment. We believe the figure briefly describes the honing process (bold text), and the right-hand side of the figure provides an example. The figure complements the explanation in the main text.

To make more clear that you're not interested in memory, you could note early on that in your studies the content is available to the participants, so there is no memory aspect (and emphasize that point in the procedure.)

Noted, on page 5, underneath the figure, we mention that this work is not focused on memory. And on page 12, we emphasize that participants will receive a summary of their discovery—to ensure that they do not forget the contents, given that the present research is not about memory.

Why did Replication 1 use 6 scenarios and Replication 2 use 5?

Replication 1 includes 6 scenarios rather than 5 because Replication 1 is a within-subjects design, and we wanted participants to receive an equal number of high- and low-specificity questions.