Review by Elizabeth Lapidow, 04 Aug 2024 17:02

Thank you for the opportunity to review this registered report. The topic is both interesting and timely and has come together quite well. My main concern at this stage concerns clarity and cohesion in how the major concepts (scientific reasoning, uncertainty, and metacognition) and, more importantly, the relationships between them are presented. There are several places where the authors' thought process in connecting these ideas is very difficult to follow. However, I expect it can all be addressed in revision, and hope my comments on specific points of confusion will be helpful.

Major Comments:

104-106 — The claim made in this sentence is sensible, but it is not well matched to or explained by the preceding paragraph. In fact, at a superficial level, it almost suggests the opposite, that how children explore uncertainty is unrelated to their ability to explicitly report it. Possibly, an explanation of the distinction the authors seem to be drawing between "ability to reflect on uncertainty" and "explicit awareness of uncertainty" would help the flow here.

This has now been clarified on page 4, line 107. Although, to keep the edits to a minimum, we focus the clarification on this paragraph.

127-128 — Echoing my previous comment, while the literature presented in the introduction clearly speaks to the claim the authors are making, they are not direct evidence for metacognition playing a role in information search — thus, the authors need to explain the logic of the connection.

True, the literature has not shown direct evidence for the connection between metacognition and information seeking, it has rather provided data to suggest such a connection by showing that improvements in children`s metacognitive reasoning (about their beliefs, uncertainty, how to seek specific forms of evidence, as well as the link between beliefs and evidence) is parallel to improvements in the efficiency of their search for supporting or clarifying evidence. We include a brief clarification of this on page 5, line 140, keeping edits to the introduction to a minimum.

160 — While I agree intuitively with all the statements in this paragraph, I am really struggling to follow the argument as it is currently presented:
-"The ability to accurately report feelings of uncertainty" would seem to refer to a performance limitation, making an accurate explicit response. However, the next sentence suggests that in fact (1) refers to children's feelings of uncertainty, while the sentence after that goes back to "ability to report." The authors should take time to address how the difference between performance and competence is relevant to their ideas.

In line with the edits made earlier in the intro, we have now made brief clarifications on page 6, line 183.

- "Identifying the most likely explanation for why the claim is wrong" seems oddly unmetacognitive, given the focus of the paper?

This is clarified on page 7, line 200.

- The prediction "this effect holds when also controlling for the ability to identify an efficient test" has not been set up by the preceding text, so its importance is unclear.

We added this tweak to improve alignment with the predictions presented in the `Proforma study design template`. However, following this comment as well as the remarks made by Reviewer 2, we realize that this was premature and have now reverted to the initial presentation of the predictions in line with the Stage 1 acceptance of the manuscript.

207— Perhaps I am simply forgetting something from previous stages of review, but why isn't there an analysis to see if whether or not children provided a plausible reason for skepticism was related to whether they provided a targeted empirical test?

This would be a relevant next step indeed. However, given that the current study only invited half of the children to reflect on their relative uncertainty about the claim, and considering that we expected only some of these to test the claim, we did not formulate an overarching hypothesis about the link between reasoning and test efficiency within that group as the N would be low. Post-hoc observational analyses in response to this comment suggests that children who do not provide a plausible reason for their uncertainty are more likely to suggest an inefficient (Total N = 16) as opposed to efficient test (Total N = 3). Similarly, children who provide a plausible reason for their uncertainty appear more likely to suggest an inefficient test (Total N = 34) as opposed to an efficient test (Total N = 21), and more likely to suggest an efficient test compared to children who do not provide a plausible reason. However, these observations are not supported by statistical chi square tests (all p`s > .05). What we do find though, and as reported in the manuscript, children who provide a plausible reason for their uncertainty appear more likely to want to test the claim compared to children who provide no plausible reason.

227 — The "Select question" also helps to control for age-related differences in children's ability to respond to a generation question, regardless of their scientific reasoning. This strength could also be mentioned here.

Good point! This note has now been included on page 9, line 258.

300 — Perhaps I have missed something — but I don't see why "the yellow thing is pink" is "something to be less sure of"? Given that it is objectively false, shouldn't children be 'sure' it is 'not true'?

The familiarization task was intended to introduce children to the basic two-step set-up of the task in which they are first asked a belief and then an uncertainty question. This would allow children to identify the selected statements as either true or false, without triggering uncertainty about what to belief. Formulating the claims during this phase as less clearly true/false would have preempted the experimental task.

348 — Given the level of detail with which the rest of the presented, I think it would be consistent and helpful to include one example of the three options on the Selection Task in this section (or potentially as an additional figure).

An example trial is now included on page 13, line 391.

373 — I am surprised that so few children (less than 30%) indicated they wanted to find out if the claim was true, given that these claims were selected to be surprising. Given that the design question was only asked for children who said yes, that means that the N's in several analyses were much lower than the target 175. Can the authors comment on this and the influence on the interpretability of their results.

We realize that this came across slightly unclear, and have now included a clarification on page 16, line 416. In the total sample, roughly 74% (N = 131) of the children wanted to find out if the claim was true or not. This was slightly less than we initially anticipated (ca. 80%). However, our primary concern for the analyses was not the slightly lower N, but that we had not taken into account that some of the children who wanted to find out the truth would not necessarily verbalize any response, making the planned coding very restricting and leading us to explore alternative ways of coding (as reported in the manuscript). The 29% referred to in the Coding-section reflects the subsample of children within in the Prompted condition who wanted to find

out if the claim was true or not (the proportion of testers within the Prompted condition was 57%). We reported it this way to illustrate the proportion of children we are running the reliability estimate on.

375 — For the sake of clarity (even at the expense of brevity) it might be a good idea to replace "T1" T3" and so on, with "Experimental Test Question" "Selection Task Question 2" and so on.

We realize that the use of abbreviation terminology could negatively affect readability, however so might long labels. We have thus left the abbreviated trial labels as they were, but include a clarification of what the T`s refer to on page 16, line 418.

410-422 — Did the trials differ in how children's responses would relate to this coding? For example, in the trial described earlier, all three objects need to be lifted to assess the claim that "the small one is the heaviest" — and there wouldn't seem to be a way of proposing "too much exploration" as it is explained here.

Correct. Only trials were only two of the three objects had to be manipulated to test the claim could lead to the conclusion that too much exploration was suggested. This was included as a control to ensure that not all claims could be solved by lifting all objects, and that children were required to construct nuanced assessments fitting to each claim.

Results & Discussion — I am unsure of what to make of the fact that younger children were less skeptical of the surprising claims than older children, given that the claims were selected to be surprising. I am very open to hearing the argument against this concern: but I worry it undermines the interpretation of children's later responses as a measure of how testing uncertain claims changes with age if there's evidence that different ages were not experiencing equal uncertainty about those claims (and even more so in whether they saw them as true or false).

Indeed, younger children's limited testing of surprising claims in prior work could be due to their not being surprised. However, we believe that is not the case here as 74% of the children wanted to test the claim. That said, to limit the chances that this would affect our analyses, a key element of the current study was not the level of surprise per se, but whether, if uncertain about what to believe, children would want to resolve their uncertainty.

594 — What is the 74% capturing? How does it differ from the 30% reported in 373?

Apologies for the confusion, we have now included a brief clarification on page 27, line 646: Approximately 74% of the total sample wanted to find out if the claim was true or not (we had expected this to be ca. 80%). The 29% referred to in the Coding-section on how we estimated reliability for the Reasoning question reflects the percentage of children in the Prompted condition who wanted to find out if the claim was true or not (i.e., roughly half of the of testers in the total sample → 29%).

689 — The distinction between "understanding why you are skeptical" and "skills to reason scientifically" strikes me as odd. There are strong proponents (D. Kuhn, for example) of the idea that mature scientific reasoning -requires- mature metacognition, for exactly this reason. What alternative to that are the authors proposing?

In addition to the clarifications made in the introduction, we now also include explicit examples to clarify this distinction on page 31, line 741.

762-764— Since this section and previous sections of the Discussion do not include results on children's ability to seek out evidence, this statement is premature here.

Agree, we realize this was a bit premature and have edited page 36, line 825.

774-778 — Perhaps I have significantly misunderstood something, but isn't improvement in scientific reasoning what the authors are claiming the difference between older and younger children's performance is capturing?

We hope our edits have helped clarify this issue. Our key prediction is that scientific reasoning and knowledge of how to efficiently test a claim is not sufficient to prompt actual testing of a claim. Rather, a key pre-requisite to such testing is recognizing the *cause* of ones uncertainty and the *necessity* of such testing if wanting to resolve that uncertainty.

812-832 — I think the claims made in this paragraph are impressive and plausible, however, how does this square with the findings (for example from Lapidow et al 2022) that children's exploration of uncertainty does not rely on their ability to reflect on their uncertainty?

Indeed, children may engage in exploratory behaviors without relying on an ability to reflect on the uncertainty driving this behavior. However, our claim in this paper and paragraph is that children are unlikely to spontaneously test claims *efficiently* if they are unable to reflect on the cause of their uncertainty about a claim. That is, we might see that children want to explore, but that without a clear rationale for their exploration it will be less targeted. In other words, our claim is not about whether exploration happens but what kind of exploration happens. We include a reflection on this point on page 39, line 886.

Furthermore, how do we separate this from a generalized "ability to articulate"? That is, if I don't know how to verbally express "I have uncertainty about what's in the box because I haven't seen what's inside it," isn't it very likely that I also don't know how to verbally express "I would look inside the box" in response to the experiment question?

True. A limited capacity to formulate a response could be a common challenge of both expressing the cause of one's uncertainty and expressing the desired/efficient testing strategy. Indeed, the fact that several children said that they wanted to find out the truth of the claim, but then didn`t provide an explicit suggestion as to how they would do it could speak to such an issue. However, we find this unlikely, given that our most restricted analyses where these children are excluded revealed a clear link between children`s capacity to provide a plausible reason for their uncertainty and their inclination to suggest an efficient test.

The "Select question" also helps to control for age-related differences in children's ability to respond to a generation question, regardless of their scientific reasoning. This strength could also be mentioned here.

Minor Comments:

79-82 — This sentence is oddly worded and needs to be simplified.

This has now been revised on page 3, line 81.

150 — What were the ages of the children/age differences in Cottrell et al., 2022?

Information about age has now been included in the revised manuscript on page 6, line 170.

289 — I expect this will be made clear below, but could the authors include whether or not the prompt, belief question, etc were also included on the Select Task trials?

In the Selection task children were only asked what they would do to find out whether the claim was true or not. This information is included in the manuscript on page 11, line 321: "…but rather than asking children whether they want to find out the truth of the claim and if so how, they were simply asked to select between a set of three…".

Review by Amy Masnick, 19 Jul 2024 17:35

This is a Stage 2 review of the manuscript Knowing why: Children's reflection on their own uncertainty about an adult's surprising claim increases their tendency to efficiently test that claim.

The study appears to have been conducted as planned, with a total of 174 children in the sample, as predicted. The Introduction is the same, and the hypotheses proposed were followed, with a tweak to the third hypothesis that with increasing age, children will be more likely to suggested targeted empirical tests for a claim. The original prediction did not qualify this prediction but the update added "and that this effect holds also when controlling for their ability to identify an efficient test when provided with multiple options." The fact that this tweak was added at this stage may need to be added to the paper.

We apologize for not being clear on this point. The tweaking of Hypothesis 3 intended as a clarification and to align with the `Proforma study template`, as we realized that the manuscript only implicitly referred to this prediction as part of prediction 5. However, based on this comment and the comment from Review 1 that this tweak is premature at that stage of the text, we have reverted to the original write-up as accepted at Stage 1.

The sample is clearly described. It was a group of children from generally high socioeconomic families, with the overwhelming majority (~83%) coming from a family where at least one parent with a Bachelor's degree, and about 75% from families with income above the national Norwegian median.

Data were largely coded as described, with high agreement and Cohen's K. Because of a large number of children who did not directly suggest testing the claim but indicated interest in exploring it, a few new variables were coded for exploratory analyses, categorizing children by whether they wanted to test the claim, whether they proposed efficient tests or not, or never wanted to explore. Only a handful of children were excluded from the study, and the reasons for these exclusions are detailed.

In the Results, the authors are clear about which analyses were planned and which are exploratory. I think it might be helpful to clarify what is meant by Block, as that term is only used in the Results section right now.

We now include a clarification on this point in the Methods section on page 13, line 374.

In testing Hypothesis 1, the effect of age was actually the reverse effect predicted, and exploratory analyses detailed that changing the order of introduction of variables into the model led to an Age x Belief interaction, such that older children are more uncertain when believe a claim and more certain when rejecting a claim, and younger children show the opposite pattern.

In Hypothesis 3, the original proposed outcome variable, the total number of times a child suggested an efficient test strategy, led to a model that supported the prediction of the ability to target empirical tests increasing with age. However, as noted above, this is where additional outcome variables were created and tested, to allow for the inclusions of a larger subset of the participants. Exploratory analyses note that the same pattern held even when using the newly-created variables.

See response above.

Hypothesis 4 was not supported as planned regarding the effect of prompting, though exploratory analyses provided partial support.

The exploratory analyses are clearly identified as such, and seem reasonable given the restrictions of the proposed coding scheme. They also allow for more nuanced understanding of the data.

Table 1 adds a column to note which of the confirmatory hypotheses were supported, and which were not, and summarizes some of the exploratory analyses, for a clear overview.

The Discussion clearly walks through the findings and offers solid explanations for the findings. The final conclusions do not differentiate the preregistered analyses from the exploratory ones, though the rest of the paper, from the Abstract, to the Results, to the main part of the Discussion, makes this point clearly.

We have now revised the conclusion to clarify that the results reported refer to both planned and exploratory analyses on page 40, line 1003.

Overall, I think the study is well-done, in accordance with the preregistered plan, and the analyses and findings are clearly described.