Reviews

Reviewed by Amy Masnick, 30 Oct 2022 20:22

I thank the authors for the substantial edits made to the paper, and I think the design and background are much more clearly justified now. I have a few comments on some of the specifics below.

First line: "Across domains, older children appear to be more intentional in their actions…" How old are older children here?

We have now reformulated this first sentence to better illustrate that we first talk about developmental change more generally, before we in the next sentence include a specification of the age range of interest to the current investigation (p. 2).

Minor point: In the Abstract, I assume it is meant, assigned to one of two conditions?

Correct, this has now been revised (p. 1).

p. 6 lines 214-215. Why is the effect of prompting expected to be stronger among younger children? Is that simply because the older children are expected to be more likely to generate an efficient test without prompting so there is less room for improvement? Or is there something beyond that?

One of our main hypotheses of this paper is that older children's inclination to test surprising claims is rooted, amongst other, in an understanding of *why* they became uncertain about the claim, which then enables them to more easily delineate the relevant test to dispel this uncertainty. Younger children, on the other hand, are expected to be less explicitly aware of the reasons for their uncertainty, and thus less likely to follow up on their uncertainty with an efficient test. Therefore, when prompting children to reason about the relative uncertainty created by the claim, we believe there is more room for improvement in younger compared to older children.

Line 224 As I read this, I was curious about the operationalization of "plausible explanation" but I know that is discussed further later.

To clarify this point, we now include a brief operationalization also on page 6.

Line 227 Explore question – I assume there is an expectation that the vast majority would want to explore? Is it a problem if too many children say no? Is there a reason you are not just asking directly how they would test if they wanted to find out the truth of the claim? That is basically the same question asked in the second task, but with an opportunity for an open-ended response rather than a choice among fixed options. I also note you have no research questions explicitly looking at the responses to the Explore question, just the follow-up Design question.

Based on prior work we expect that a majority of the children will suggest to explore the objects, and that the main distinction lies in the manner in which they suggest to do so. To avoid a confound between the interest in acquiring further information and the

ability to do so efficiently, we split the task in two using the Exploration and the Design question. As is pointed out though, there is the risk that only a limited number of children suggest to explore leaving few children to receive the Design-question. However, while we don't believe that this is likely, a scenario where few children explore would also be informative as it questions children's motivation to engage in exploration in prior studies.

In light of this comment though, we have made some minor changes to the analytical plans for Hypotheses 3 and 4. We noticed that by relying on an average score when analyzing the efficiency of children's testing, children who suggest to explore only once but who that one time suggest an efficient test will come across as more likely to test a claim efficiently, while a child who always suggests to explore, and does so efficiently three out of four times will come across as less likely to engage in efficient testing. To better convey such nuances in the data, we now include first a frequency-based analyses of children's tendency to efficiently test the claim (i.e., counting the number of times they suggest to test the claim efficiently), before we use the average score to analyze the stability in children's tendency to efficiently test the claim (see the proforma study design template and the revised R-script).

Line 309 Prediction 4 also predicts an age effect/interaction.

See mentioned changes above.

Lines 319-328 I like the addition of the separate test of three surprising claims with multiple choice options to confirm children of all ages can accurately assess testing strategies.

Great!

Lines 406 "how certain or uncertain they were in the belief" – it is a dichotomous question of sure or not sure, not a rating, correct?

Correct, this has now been revised.

Lines 679 – would a mechanistic but physically inaccurate answer count as plausible?

We were unsure about what is meant by "mechanistic by physically inaccurate". If "physically inaccurate" means unrelated to the target object, then the answer is no. In order to count as plausible, the explanation has to focus on the target object. If by "physically inaccurate" means that while the mechanism described by the child exists (differences in density) but cannot apply to the objects presented to the child (because they were told the objects were all made of the same material), then the answer is also no.

Table 1 is a helpful study template.

For Q 3 in the table, do you mean Design average is the DV? Reasoning average is only for the prompted condition. But then Q 4 analyses seem to address many of the same questions, so I am a little unclear on the distinctions here.

Correct, the dependent variable for Q3 should be related to Design and not Reasoning and is only analyzed among children in the unprompted condition. Note however, that

we now include a preliminary step to this analysis. In the revised analysis, we first plan to assess the frequency with which children suggest an efficient test (labeled as analysis I), before we investigate the stability of this measure by using the average score (labeled as analysis II). This has been updated in the revised manuscript and attached R-script.

With Q4, we assess the role of prompting children to reason about their uncertainty on their tendency to test the claim by introducing Condition as an independent variable as well as the Age BY Condition interaction term. Note that we also here now plan for a preliminary assessment of the frequency of children's tendency to suggest an efficient test before analyzing their average scores.

I wondered about pilot testing of stimuli, confirming these are often surprising, though I know you've made reference to developing your stimuli based on previous studies, so if that is the source, perhaps that could be noted explicitly.

The stimuli were initially developed based on a combination of prior work, of which we have now included a reference to on page 12.

Overall, though, I am much more supportive of the current iteration of the proposal and think it has the potential to add to our understanding of children's developing reasoning and metacognitive skills.