Dear Dr. Heyman,

We thank you and the reviewers for your extremely helpful comments on our submission called “Don’t throw the associative baby out with the Bayesian bathwater: Children are more associative when reasoning retrospectively under information processing demads”. We believe that the manuscript has improved substantially based on these comments. In what follows, we go through each comment made and show precisely how we have addressed those concerns in our revised manuscript. Our responses are written in red. The original comment is **in white**.

**Response to Reviewers**

**Reviewer 2.**

My primary concern with the overall manuscript is that there is a mismatch between the motivation of the study in the Introduction and the overall conclusions drawn in the General Discussion. This makes it difficult to assess whether the study is well-motivated how to best assess the researchers interpretations, as well as the overall contributions of this manuscript. On the one hand, the researchers motivate the study primarily as using backwards blocking paradigms to distinguish between Bayesian and Associative approaches for causal learning. On the other hand, the General Discussion focuses more directly on how the computational models fit the behavioral data and begins to explain more clearly how the models relate to the specific conditions and thus relate to the Bayesian vs. associative debate. Given the focus seems to be related more directly to the model comparisons, the researchers should revise the manuscript so that it is motivated more clearly by the computational modeling approach.

We thank the reviewer for this extremely important comment. We have now revised the Introduction to address mismatch problem that the reviewer (rightly!) raised. In particular, we now include a discussion of how previous work on retrospective reevaluation in children is limited because few, if any, attempts have been made to fit associative and Bayesian models to children’s data. We argue that this is important because it enables one to quantify precisely how much children’s retrospective reevaluations reflect associative processes or processes that are based on Bayesian inference. Such models also make it possible to determine whether different models provide better accounts of different aspects of the same data – this can reveal the use of multiple cognitive mechanisms in a single task. This discussion can be found on pages 4 and 5.

I am unclear on whether the researchers are making any specific developmental claims based on this data. On p. 6, they reference several studies that claim that younger children rely on associative strategies or that children fall back to associative strategies when under higher cognitive load. This overall argument is also common across this literature. However, in this particular study, the researchers do not present any clear developmental differences. Although they do not make explicit statements about how this study relates to those broader issue related to developmental change, I could see how they might want to use this data to speak to those issues. For example, are the researchers claiming that both Bayesian and Associative strategies are present early in development (infancy?) but that children may vary in which strategy they use over development and as information processing demands increase? Or, would they claim that these findings support conclusions that that associative learning strategies precede Bayesian strategies? I think the researchers should include in the General Discussion a discussion that makes these implications (or the absence of these implications) more clear.

We thank the reviewer for this comment! We have now included a discussion in the GD about precisely this issue. We note that although we failed to observe a developmental effect—children’s backwards blocking and indirect screening-off inferences were unrelated to age—the study does have developmental implications, especially to the extent that there’s a relation between a person’s age and their information-processing abilities. We specifically argue the following in the General Discussion (p. 33-34):

“A fourth potential criticism concerns the absence of developmental change in children’s retrospective reevaluations: Children’s backwards blocking and indirect screening-off inferences were unrelated to age in both experiments. Although we failed to observe an age effect, the current results do have developmental implications. If we are correct that children resort to more associative forms of processing when their information-processing capacities are stretched, then these results suggests that if younger children are tested in a replication of the current study their inferences should be even more associative than the 5- and 6-year-olds tested here. This is because younger children presumably possess less robust information-processing abilities than older children and thus should be more affected by the increase in the number of objects used (relative to past studies on retrospective reevaluation) than the 5- and 6-year-olds tested here. Conversely, if children older than that tested here or even adults are tested in a replication of the current study, then not only should they be less affected by the increase in the number of objects presumably because they possess more information-processing abilities than the children tested here, but their inferences should also better align with the predictions of the Bayesian model than the associative model.”

The researchers note that adding a third object requires reasoning over larger hypothesis space (p. 4), but this point is relatively under-developed. They do state that it “may introduce additional information processing demands” (p. 4), but they do not state how they would impact any decision between Bayesian and associative approaches. Some of this is discussed more directly in the computational modeling section. Moving some of that text to earlier in the manuscript would be helpful.

We have now clarified how reasoning about 3+ objects introduce information-processing demands that reasoning about 2 (or fewer) objects doesn’t. This text can be found on p. 5 second paragraph. To summarize, the reason reasoning about 3 or more hypotheses introduces additional information processing demands is because the underlying hypothesis space goes from four possibilities in the case of two objects (i.e., object A and B can be blickets; object A but not B is a blicket; object B but not A is a blicket; neither object A nor B is a blickets) to eight possibilities in the case of three objects. If we introduce just one more object, going from 3 to 4, then the hypothesis space grows to 16 hypotheses. If children are sensitive to this increase and possess limited information processing abilities, it may be the case that in situations such as this they resort to a simpler form of reasoning that is better captured by simple associative processes than sophisticated inferential processes like Bayesian inference.

The predictions on p. 7 are only expressed in relation to whether A does or does not make the machine go, ignoring the D block. This made it more difficult to understand the contrast between conditions and the related predictions.

We thank the reviewer for this comment and have revised this paragraph to better clarify what the relevant and specific predictions are for the backwards blocking and indirect screening-off conditions. The relevant text can now be found on p. 8 starting with, “Participants were said to engage in backwards blocking…”.

Participant demographics – p. 7, ln 52: “82% White/Caucasian (compared with 83%).” I do not understand what “compared to 83%”. Also, I think “Over 120K: 38K” should be 38%.

We have removed “(compared with 83%)” (it was included in error) and have changed 38K to 38%,

Although mixed models can handle unbalanced designs, I think it makes it difficult to interpret the results. For example, given that block D is only present in control trials, are the researchers able to include an overall analysis with an Object factor? On p. 12, ln 38 they state, “This reflected the fact that the scores for object A were collapsed over Trial Type (in which A was seen by itself during the experimental trials but in combination with other objects during the control trials).” Doesn’t this then suggest that analysis is not quite right -- that they cannot compare block A to block D?

We thank the reviewer for pointing this out. This was an error on our part – as was true for all of the objects, the scores for all objects were collapsed over Condition, not Trial Type. We removed this sentence from the text.

Did the researchers find any age effects in the follow-ups analyses for the Age x Condition x Object interaction? If not, how do they then interpret that interaction? Why is it not then only a Condition x Object interaction?

We thank the reviewer for this question. Based on Reviewer 3’s concerns, we have re-run the analyses by fitting more appropriate statistical models—namely, mixed-effect logistic regression models—to the data. As was true before, there was no age effect when these new analyses were fun. It was also **not** the case that the Age x Condition x Object interaction was significant. Nonetheless, we discuss in the GD how are results nonetheless have developmental implications. That text can be found above in our response to Reviewer 2’s second comment.

Given that the objects are different (at least from the child’s perspective) across trials, is it meaningful to make comparisons between objects across trials? Block A and Block D might be relevant across trial number (although that would be matched to whether the backwards blocking or indirect screening off trial), but Block B and Block C are only different based on spatial position.

This is an important comment. We agree with the reviewer that the objects, from the child’s perspective, are physically different across trials (i.e., they’re different colors across the experimental and control trials). However, despite these superficial surface differences, any two pairs of objects can have *functionally* equivalent roles across trials (i.e., between the experimental and control trials). For example, objects A and D, although physically different, serve in functionally the same role: Both objects participate on the machine alone. Likewise, although B and C are physically different across trials, they are functionally the same: Neither B nor C participated on the machine between the experimental and control trials.

For the Condition x Trial type x Object interaction (p. 13), why are the researchers not testing for a Trial Type x Object interaction within each condition. They also only report Object effects, but not effects of Trial Type or any interactions.

We apologize for this oversight. We have now reported the two-way interactions for each condition. As expected, the two-way interactions for both conditions revealed a significant main effect of Objects, a main effect of Trial Type, and most importantly a significant interaction between both. This analysis can be found on p. 15.

Experiment 2, Figure 3: The x-axis labels are now ‘main’ and ‘control’ rather than ‘experimental’ and ‘control’.

We have fixed this issue.

**Reviewer 3.**

Abstract: The summary of the paper, its aims and findings, is too vague.

We have significantly revised the abstract to make it clearer what the study was about, why it is important, and what we found. Here is that text:

“Causal reasoning is a fundamental cognitive ability that enables humans to learn about the complex interactions in the world around them. However, the mechanisms that underpin causal reasoning are not well understood. For example, it remains unresolved whether children's causal inferences are best explained by Bayesian inference or associative learning. The two experiments and computational models reported here were designed to examine whether 5- and 6-year-olds will retrospectively reevaluate objects—that is, adjust their beliefs about the causal status of some objects presented at an earlier point in time based on the observed causal status of other objects presented at a later point in time—when asked to reason about 3 and 4 objects and under varying degrees of information processing demands. Additionally, the experiments and models were designed to determine whether children’s retrospective reevaluations were best explained by associative learning, Bayesian inference, or some combination of both. The results indicated that participants retrospectively reevaluated causal inferences under minimal information-processing demands (Experiment 1) but failed to do so under greater information processing demands (Experiment 2) and that their performance was better captured by an associative learning mechanism, with less support for descriptions that rely on Bayesian inference”

Introduction: Papers cited in the introduction are not listed in the reference list.

We thank the reviewer for spotting this. We’ve fixed this issue.

There is no information on whether the study had ethic approval. There is no information on whether parents gave informed consent etc.

We now indicate in the title page that the study was not preregistered but was approved by the study site’s IRB. We have also now indicated in the Procedures section on p. 10 that we obtained informed parental consent as well as child assent.

Procedure: I was surprised that 5-year-olds were required to read text during the experiment. It was unclear from the procedure whether this text was read out by anyone?

We apologize for the confusion. Neither the 5-year-olds nor the 6-year-olds were required to read any text. Instead, the experimenter read the text that was built into the video. This served to standardize what was said across participants. We have amended the text to make this point clearer. Now on p.9 it reads: “Finally, the videos contained a built-in script, which experimenters, but not the study participants, read.”

Procedure: It is not even clear whether any experimenter was present. Did an experimenter read out the questions or was this done by a caregiver?  How were children’s answers recorded?

We have now clarified in the text that participants’ responses were coded offline after each study session. The text on p. 10 now reads: “Finally, all study responses were coded offline after each study session.” In terms of whether an experimenter was present and read the script, we have made it clearer in the text that an experimenter was present throughout the entire session. The text on p.10 now reads: “Although study responses were coded offline, an experimenter was present throughout the entire study session.”

Procedure: There are problems with the references to figures and tables: On page 10 it says, “A schematic of this procedure is shown in Table 1.” – Does this refer to Figure 1? It cannot be Table 1.

The reviewer is correct. We meant to refer to Figure 1, not Table 1. We have made the corresponding change.

Sample size and pre-registration: How was the sample size determined?

We thank the reviewer for this comment. The text now reads, “Sample size was determined based on previous studies on backwards blocking reasoning in human children (e.g., Griffiths et al., 2011; Sobel et al., 2004)”

Sample size and pre-registration: Where these analyses pre-registered?

We have included a note in the title page that the study was not preregistered. However, we have included a link to our OSF which is freely available and contains the deidentified data along with the analyses.

Analyses: Was the dependent variable the number of time children gave yes responses for each trial type, with scores ranging from 0 to 2?  Or was the dependent variable a binary measure? The first sentence of the results of Study 1 states: “Figure 2 shows the number of times children responded “yes” to the question “Is this a blicket” for each object. Using this as the dependent variable …” As far as I understood, each child was only asked once for each object, which would suggest that the DV was binary.

We really appreciate the reviewer for this (and the following comments) and believe that the results section have significantly improved. This is because we’ve re-run all the analyses by fitting mixed-effects logistic models to the data, which the reviewer (rightly!) notes is the more appropriate test to run. We have chosen this kind of model because our response measure is binary. In addition, we have clarified in our figures that what is being shown are participants’ responses expressed as percentages.

Analyses: Was the model used a general linear mixed effects model (for continuous outcomes) or a generalised mixed effect model for binary outcomes (i.e. a binary logistic regression)?

We thank the reviewer for this comment and agree that given that the data were binary, it is more appropriate to fit the data with the binary mixed-effects model. As the reviewer will be pleased to see, we have significantly modified the results sections of Experiments 1 and 2 to reflect these more appropriate analyses. Crucially, and as the reviewer will also be able to see, none of the main results or conclusions change.

Analyses: Were the model assumptions met (for whichever type of model was fitted)?

We thank the reviewer for this comment. Yes, the model assumptions (for the mixed-effects logistic regression) were met.

Analyses: The fact that trial number is included as a predictor suggests that the DV is binary. Much more details are needed. Did the researchers fit a GLMM using R? If so, which interaction terms were included?

See above.

Analyses: What about relative model complexity (the relation between the number of predictors and sample size)?

The number of predictors in the main analyses with 2 predictors is considerably less than the number of participants (32 in Experiment 1 and 33 in Experiment 3). Therefore, we are not concerned that the models have overfit the data. Nonetheless, we thank the reviewer for this comment.

Analyses: Analyses of study 2 – I was surprised that the authors report t-values (e.g., page 18) given that they either have binary or ranked data as dependent variables.

We have completely revised the analyses of Experiment 2 to reflect the more appropriate analysis, which involves fitting binary mixed-effects logistic models to the data.

Analyses: It is not easy to follow how the analyses relate to the predictions made. It is confusing to follow the predictions and their tests, because way the conditions and trial types are described varies in each section. The use of terminology should be made be consistent and clear throughout.

We hope that the new analyses we report will help to clarify this (justified) concern. We have also made the terminology and consistent and clear throughout the ms.

Analyses: A table should be included to clarify the condition names, what children witnessed in the first and second demonstration (see e.g., McCormack et al. 2009, Table 1) for each of the two experiments.

We have now included 2 tables—one to show the task structure in Experiment 1 and the other to show the task structure in Experiment 2.

Discussion: Experiment 1 – more information for the rational of Experiment 2 is needed.

We have better justified the rationale for Experiment 2. The reviewer can find this text on pages 16 and 17.

Computational models: I could not follow the rational for how the models were set up and how their fit was evaluated and compared. For example, it seems that the authors calculated an average of the RMSE and MAE and then, based on simple visual comparisons of the numerical values rounded to two decimal points, decided that the model with the lower value was the model which fit the data better than the other model.  At this, as well as other points, clearer explanations are needed regarding what was done and why.

We thank the reviewer for this comment and are happy to clarify. This is the standard approach for model comparison. In particular, the model that produces lower RMSE and MAE values is taken to be the ‘winner’. Given that any one model will only produce a single RMSE and MAE value, it is not possible to use traditional inferential statistics, such as *t*-tests, to determine whether one RMSE is statistically significantly different than the other, which we believe is what the reviewer may be looking for here.