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.

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. 4 starting with, “Second, it remains …”. 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. We did not find any age effects or interaction between age and object in follow-up analyses for both conditions separately. We have now included a footnote on p. 12 to indicate that although the three-way interaction was statistically significant, when analyzing the effect of age on choice within the backwards blocking and indirect screening-off conditions separately, that effect did not differ for the different objects. We suspect that what is going on here (and this is what the footnote is about in the manuscript) is that because at least one of the differences in slopes (the difference in slopes between objects A and C) approached statistical significance (*p* = .07) in the backwards blocking condition, whereas none of the slope differences even approached statistical significance in the indirect screening-off condition, this led to a significant three-way interaction between Age, Condition, and Object. We thank the reviewer for this comment and we hope that we have clarified the nature of that three-way interaction.

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.

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.

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 partiicpants’ 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. As we mention in the text on page 8, “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.

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.

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)?

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

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?

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

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.  
Relation between predictions and results

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.

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.

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

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.