Is Bayesian inference even necessary? Revisiting backwards-blocking reasoning in human children

Deon T. Benton1, David Kamper2, Rebecca M. Beaton1, David M. Sobel2

1Vanderbilt University  
2Brown University

Address correspondence to Deon T. Benton, Department of Psychology and Human Development, Vanderbilt University, Peabody College, 230 Appleton Place #552, Nashville, TN 37235

Data availability statement: The code and network-modeling scripts are available upon reasonable request.

Conflicts of interests: none

Abstract

Causal reasoning is a fundamental cognitive ability that enables humans to learn about the complex interactions in the world around them. However, the available evidence suggests that the mechanism or set of mechanisms that underpin causal reasoning are not well understood. It is unclear, for example, whether causal reasoning is underpinned by a Bayesian mechanism, an associative mechanism, or both. Some theorists have argued that a Bayesian mechanism underpins causal reasoning because it can better account for backward-blocking (BB) and indirect screening-off (IS) findings in children and adults (e.g., Sobel, Tenenbaum, & Gopnik, 2004). However, the evidence is mixed about the extent to which learners engage in both kinds of reasoning. Here, we report three experiments that examine to what extent adults engage in BB and IS reasoning using the blicket-detector design (e.g., Gopnik et al., 2001), what mechanism best explains their behavior in this task, and under what conditions are adults’ causal ratings consistent with the predictions of the three competing computational and analytical models. The results of Experiment 1 revealed that adults’ causal ratings in the backwards-blocking condition (as well as in the indirect screening-off condition) were consistent with the predictions of the traditional and modified Rescorla-Wagner models when asked to reason about two objects. The results of the present study suggest that adults use associative processes to reason about two objects but a Bayesian-inference-like process to reason about three or more objects.

Keywords: causal reasoning; causal mechanisms; computational models; analytical models; associative learning; Bayesian inference

There is perhaps no ability that is more central for learning about how the world works than causal reasoning or the capacity to reason about cause-and-effect relations. This is a key cognitive ability because it enables human learners to encode causal relations to inform prediction and inference (e.g., Oakes & Cohen, 1990; Rakison, Smith, & Ali, 2016; Schlottmann & Shanks, 1992), to intervene on those relations to generate new effects (e.g., Gopnik et al., 2001), and counterfactually to reason about causal events to determine what would have happened if the chosen intervention had not been undertaken (e.g., Harris, German, & Mills, 1996; Sobel, 2004).

Despite consensus among researchers about the importance of causal reasoning, there is much less consensus among theorists about the cognitive mechanism that underlies this capacity. For example, it is unresolved whether domain-general mechanisms such as associative learning underpins causal reasoning or whether—as has recently been suggested by some theorists (e.g., Gopnik et al., 2004; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2017)—causal reasoning is grounded in a Bayesian-inference mechanism. One empirical finding about which domain-specific and domain-general theorists have disagreed considerably concerns whether an associative-learning mechanism or a Bayesian-inference mechanism subserves human beings’ capacity to engage in one form of retrospective reevaluation called backwards-blocking reasoning. This form of reasoning involves learning blocking or discounting redundant causal cues when other cues are shown unambiguously and in isolation to produce effects (e.g., Blaser, Couvillon, & Bitterman, 2004; Shanks, 1985; Shanks & Dickinson, 1987; Sobel et al., 2004). The aim of the experiments reported here was twofold. First, it was designed to examine whether and to what extent human children engage in backwards-blocking reasoning in a new context. Specifically, in contrast to previous studies on backwards-blocking reasoning in human children that has tended to ask children to reason about two objects, here we examined whether children could engage in this form of reasoning when asked to reason about multiple objects. Second, this study was designed to illuminate whether an associative-learning mechanism or a Bayesian-inference mechanism underlies children’s backwards-blocking reasoning performance in the current situation.

**The** **emergence of BB reasoning**

The ability to reason about causal events is thought generally to emerge between 18 months and 5 years of age (e.g., Benton, Rakison, & Sobel, 2021; Gopnik & Sobel, 2000; Gopnik et al., 2001; Kimura & Gopnik, 2019; Meltzoff, Waismeyer, & Gopnik, 2012; Sobel & Kirkham, 2006, 2007; Sobel & Munro, 2006; Walker & Gopnik, 2014; cf. Sobel & Kirkham, 2005). Although researchers have used a variety of paradigms to examine causal reasoning in human children (for a review see Bullock, Gelman, & Baillargeon, 1982), here we focus on research that has used the blicket-detector design. We focus on this paradigm for three reasons. First, it has been used most extensively to test children’s causal-reasoning abilities as well as to assess their ability to engage in backwards-blocking reasoning. Second, versions of this design have been used to evaluate adults’ causal reasoning abilities, which may support cross-study and between-age comparisons. Third, we focus on this paradigm because the notion that human reasoners use Bayesian inference to reason about causal events was introduced within the context of the blicket-detector studies and in concert with key advances in computer science, philosophy, machine learning, and statistics (for a review, see Gopnik et al., 2004).

In studies that use this design, children are introduced to a machine called the "blicket detector" and told that the machine lights up and plays music when certain objects—namely, "blickets"—are placed on it but not when other objects are placed on it. Following a series of events in which the detector activates (or not), children are then asked to determine which objects are blickets and to “make the machine go” by placing the blicket on the machine. Of the findings that have been reported by researchers who have used the blicket-detector methodology, perhaps none have been more controversial than that by Sobel, Tenenbaum, and Gopnik (2004). They showed that 4-year-old children—and to a lesser extent 3-year-old children—can engage in BB reasoning and IS reasoning. Children were shown initially that two novel objects A and B together caused the detector to activate and then that object A alone either failed to activate the detector (i.e., AB+ A-; IS condition) or activated the detector when placed on it (i.e., AB+, A+; BB condition). Children in both conditions were then asked which of the two objects were blickets and to make the machine go by placing the blicket on the detector. It is worth noting that the BB condition is so called because after observing that A alone can activate the detector, children who engage in this form of reasoning are thought to disregard or block retrospectively object B as a potential cause because A was shown unequivocally to produce the effect. In contrast, the ISO condition is so called because B is assumed indirectly to "screen off" or to block object A as a potential cause given that A alone failed to activate the machine.

Sobel et al. (2004) found that when children were subsequently asked to make the machine go, the 4-year-olds during the ISO trial—and to a lesser extent the 3-year-olds during the same trial—responded by placing object B on the machine. In contrast, during the BB trial these same children responded by placing object A on the machine. Subsequent research by Sobel and Munro (2009) found that 3-year-olds, like the 4-year-olds in Sobel et al. (2004), could engage in BB and ISO reasoning if the activation of the detector represented desires rather than a physical effect: the 3-year-olds categorized object B as a blicket in the ISO condition but were less likely to do so in the BB condition but only when the machine was called “Mr. Blicket” and said to like blicket objects. These findings have since been interpreted not only as evidence that human children can engage in backwards-blocking reasoning but as evidence that this form of reasoning is underpinned by a Bayesian-inference mechanism rather than by an associative-learning mechanism. The crux of the Bayesian-inference account is that human learners use a simple form of Bayes’ rule to reason about causal events and to choose the causal hypothesis—within a space of hypotheses that is potentially super-exponentially large—that is most consistent with the observed data (e.g., Sobel et al., 2004; Gopnik & Wellman, 2012). More specifically, this process involves combining prior beliefs about each hypothesis with observed data to update the (posterior) probabilities of each of the hypotheses in the psychological hypothesis space.

One specific kind of associative-learning model that has come under scrutiny is the traditional Rescorla-Wagner (henceforth, RW) model (e.g., Rescorla & Wagner, 1972). The previous findings challenge the RW model for three key reasons. First, this model predicts that B should be treated equivalently across the BB and ISO conditions, which is a prediction that does not accord with participants’ actual performance. There are two reasons for this. First, the association between object B and the outcome was identical in both conditions; that is, B was shown to produce the effect (in combination with object A) twice in both conditions. Second, this model only makes weighted adjustments to cues that are present, which B was not during the "A" phases in both the BB and ISO conditions. This means that because object B is absent during the A phases of the BB and ISO tasks, the RW model predicts that the associative strength between object B and the blicket effect should remain unchanged across the experimental trials in both conditions, and thus further predicts that participants should treat B equivalently across both conditions. Second, the RW model requires many learning trials for reliable associations to be established and used to make causal inferences. In contrast, in the studies cited above participants engaged in BB (and ISO) reasoning based on a handful of learning trials. Finally, the BB and ISO findings challenge the RW model because it does not naturally encode base rates and thus is ill-equipped to explain children’s sensitivity to base-rate information.

**How *should* BB reasoning be operationalized and an open question**

To date, researchers have operationalized BB reasoning in terms of the difference in the treatment of object B between the BB and ISO conditions (although see De Houwer, Beckers, & Glautier, 2002; Larkin, Aitken, & Dickinson, 1998; Griffiths et al., 2011; Kruschke & Blair, 2000; Lovibond et al., 2003; Shanks, 1985; Van Hamme and Wasserman, 1994); that is, some researchers have operationally defined BB reasoning as greater B choices in the ISO condition than in the BB condition. This way of operationally defining BB reasoning is presumably motivated by two factors. First, if the causal status of object A—which can be determined unequivocally when object A is placed alone on the machine—causes participants retrospectively to reevaluate the causal status of object B, then participants should consider B to be less of a blicket in the BB condition than in the ISO condition. This is because A by itself fails to produce the effect in the ISO condition, whereas it produces the effect by itself in the BB condition. Second, proponents of this operationalization of BB reasoning have used the fact that participants do treat object B differently between the BB and ISO condition as evidence against the RW model. Recall that the RW model predicts that participants should treat object B equivalently between the ISO and BB conditions because the frequency with which object B was paired with the machine’s activation is identical between both conditions.

To our knowledge, this operationalization of BB reasoning has not been challenged in the literature despite important shortcomings. One major limitation concerns the fact that BB is operationalized in terms of the difference in treatment of object B across the BB and ISO conditions. The problem here is that it is logically possible that participants treated object B differently between the BB and ISO conditions because they observed a positive effect during the elemental (i.e., A+) phase in the BB condition but a negative effect during the elemental (i.e., A-) phase. This would mean that participants’ treatment of object B could have been based on the fact that the two conditions differed in terms of their low-level features rather than on a true retrospective reevaluation of object B based on A’s relationship to object B during the compound (i.e., the AB+) phases in both conditions and its subsequent effect on the machine during the elemental phases. Given this limitation, we argue that a more (construct) valid operationalization of BB reasoning is to compare treatment of object B following an AB+ A+ sequence of events (i.e., the BB experimental condition) to the treatment of object B following an AB+ C+ sequences of events (i.e., the BB control condition). These two conditions differ in terms of the object that is shown during the elemental phase and its relationship to B. In the BB experimental condition, a dependency is established between objects A and B because they were shown together during the compound phase of the condition. This means that A’s causal status that is established during the subsequent elemental phase should affect participants’ treatment of B. In contrast, in the BB control condition, object C was never shown with object B, which means that C’s causal status should have no bearing on whether participants consider object B to be a blicket. Crucially, the blicket effect itself is held constant such that participants observe blicket-detector activation in both conditions. If participants engage in BB reasoning in this context, then this would provide stronger evidence that participants have access to such a mechanism, assuming that BB reasoning is treated as an indirect measure of Bayesian inference.

This limitation notwithstanding, it remains unanswered whether participants can engage in BB reasoning when asked to reason about more than two objects. This is because most, if not all, of the studies on BB reasoning in human children have tended to use two objects; that is, participants are shown an AB+ A+ sequence of events and then asked whether each object is a blicket. Nonetheless, this is an important question to answer because if Bayesian inference underlies human causal reasoning, and it is further assumed that BB reasoning is an indirect measure of the operation of such a mechanism, then it is crucial to show that participants continue to engage in BB reasoning even when they are asked to reason about three (or more) objects. This is also an important question because it is possible that when children’s information-processing abilities are taxed—which could happen when they are asked to reason about three (or more) objects—they may resort to simpler modes of causal reasoning. One such simpler mode of reasoning that participants may resort to when asked to consider multiple candidate causes is the traditional RW model. If participants’ BB performance adheres to the predictions of the traditional RW model in a multiple-candidate-cause setting, this would suggest that it was premature to conclude that the traditional RW model is an inadequate model of causal reasoning in human children. Thus, by understanding whether participants engage in BB reasoning in a multiple-candidate-cause setting, we can gain insight into the cognitive mechanism by which human children reason about causal events.

**The present investigation**

The present investigation is designed to determine whether participants’ BB performance in the context of three and four objects is better explained by a Bayesian-inference mechanism or or the traditional RW associative-learning model. Given this goal, it is important to determine what predictions both processes make for the present conditions as well as whether the predictions that these models make differ.

A **Bayesian model**

A fundamental assumption of the Bayesian inference approach is that causal induction is a process that involves representing the entire space of candidate causal hypotheses—which are expressed as parameterized causal graphical models with nodes that are connected by edges that encode the Markov condition—and then choosing the hypothesis that is most consistent with the data by application of Bayes' rule. Formally, it is assumed that, at the beginning of a task, experiment, or learning episode, an ideal Bayesian learner represents all possible candidate hypotheses, *H*, whereby each hypothesis, *h* ∈ *H*, is assigned some prior probability, *p*(*h*). This prior probability represents the learners’ confidence that the observed data were generated by a given causal hypothesis. Following observations of data, *d*, the learner then uses Bayes' rule to compute and assign posterior probabilities to each hypothesis, *p*(*h*|*d*),

,

where *p*(*d*|*h*) represents the likelihood or the probability of the data *d* under a given hypothesis *h* normalized by members of the set. Because the hypotheses in the present experiments are assumed to be deterministic (i.e., objects either produce or do not produce detector activation), the likelihoods are set to 1 whenever a link (i.e., causal relation) exists in the hypothesis (Figure 1) and is consistent with the observed data; otherwise, they are set to 0.

The first step in defining a Bayesian model of the present tasks was to specify the hypothesis space *H* and the hypotheses *h* that comprise that space. This step is necessary before Bayes' rule can be used to determine the hypothesis with the largest posterior probability. Because Experiment 1 and 2 uses three objects (i.e., three candidate causes) and Experiment 3 uses four (i.e., four candidate causes), the hypothesis space for each experiment, respectively, consists of eight hypotheses for Experiments 1 and 2 and 16 hypotheses for Experiment 3 (see Figures 3, 5, and 7 above). The specific parameterization of each hypothesis in the space is specified by the activation law, which, for all three experiments, states that the blicket detector will activate if, and only if, a blicket object contacts it. The second step in defining this model is to specify the prior probabilities of each hypothesis in each of the three experiments. If we assume that the probability that a particular object is a blicket is independent of the probability that other objects are blickets, then prior probabilities for Experiments 1-3 can be found in Tables X. These prior probabilities are then used to compute posterior probabilities for each hypothesis when new data is observed according to Bayes' rule.

Because I am assuming deterministic hypotheses, whenever a link exists, the likelihood is set to 1; whenever a link does not exist the likelihood is set to 0. Once it is determined that a link exists for a particular object, we can compute the likelihood that object A, B, or C in Experiments 1 and 2 or object A, B, C, or D in Experiment 3 is a blicket by taking the product of the likelihood that a particular object activated the detector under each hypothesis and the prior probability of each hypothesis and then summing this product. To determine the probability that object B is a blicket, for example, we can compute the following equation:

,

where is 1 if a causal link between *B* and *E* existsfor a specific hypothesis *h*; otherwise, is 0.

**Experiment 1**

The goals of Experiment 1 were twofold. The first goal was to determine whether adults engage in BB and IS reasoning. The second goal was to determine to with which model—that is, a simple Bayesian model, the traditional RW model, or the modified RW model—participants BB and IS performance was consistent. Correspondence between adults' causal ratings and the predictions of either of the three models would indicate which of three mechanisms underpins adult causal reasoning. In the study, participants were introduced to a machine called the blicket detector and were told that their task was to determine which objects make the machine activate—and thus represent blickets—and which objects do not make the machine activate. Following this brief introduction phase, participants received four trials—a backwards-blocking trial, an indirect screening-off trial, a one-cause trial, and a two-cause trial—and were asked to provide three sets of causal ratings for objects A and B across each trial. In particular, participants were asked to provide causal ratings for both objects before a given trial had been demonstrated, midway through a given trial, and after a trial had been demonstrated. The ratings reflected how causal the participants thought each object was on a scale that ranged from 0 (definitely not causal) to 100 (definitely causal).

**Method**

**Participants, stimuli, and design.** Twenty-four college students were recruited for Experiment 1 and were given course credit for their participation (11 males; 13 females).

**Stimuli and Design.** The device used in Experiment 1 was similar to the blicket detector that was used in previous studies (e.g., Gopnik et al., 2001; Sobel et al., 2004). The box was 5" x 7" x 3", was made of wood that was painted black, and had a white Lucite top. The machine operated via a remote control that was attached to the end of an electric wire that was attached to the side of the box. When the button on the remote control was pressed and the object that was predetermined to be a blicket was placed on the surface of the detector, the music began to play and the lights began to flash. If an object that was not predetermined to be a blicket was placed on the detector, the button was not pressed and the lights and music did not flash or play.

In this and all subsequent experiments, adults were asked to provide three sets of causal ratings on the same 0-100 scale during the pre-, mid-, and post-rating phases for two objects (A and B). The stimuli for this experiment were eight cube and cylinder objects (four of each), each approximately 1" in diameter and of different colors. Within each trial, the objects were of different shapes and colors and the object that was designated as the blicket was counterbalanced across participants. Two unrelated objects—which were neither cubes nor cylinders—were used in the pretest phase of the experiment.

**Procedure.** Participants were tested in a quiet room on campus. At the beginning of the experiment, participants were introduced to the machine and were told it was called a blicket detector that activated only when objects that were predetermined to be blickets were placed on it. Participants were then informed that their goal was to determine which objects were blickets and which objects were not based on the pattern of activation observed during the experiment. Following this introduction, participants were given four test trials in counterbalanced order using a Latin square. The trials used here were identical to those used in previous blicket-detector studies and included the one-cause (1C), two-cause (2C), indirect screening-off (IS), and backwards-blocking (BB) trials (e.g., Gopnik et al. 2001; Sobel et al., 2004).

Similar to previous blicket-detector studies, the 1C and 2C trials served as controls to ensure that participants understood the test events. In the 1C, participants were shown two blocks (A and B) and then observed that block A caused the detector to activate, whereas block B did not. Blocks A and B were then placed together on the machine twice, which caused the detector to activate both times (i.e., A+; B-; AB++). In the 2C trial, block A activated the detector each of the three times it was placed alone on the detector, whereas object B failed to activate the detector the first time it was placed on the machine but then activated the detector the remaining two times it was placed on it (i.e., A+++; B-; B++). In the BB trial, objects A and B were placed together on the machine twice, which caused the machine to activate both times. Object A was then placed alone on the machine, which once more caused the machine to activate (i.e., AB++; A+). The IS trial was similar to the BB trial except that object A failed to cause the machine to activate when it was placed on it (i.e., AB++; A-). In each of the four trials, participants were instructed to rate on a scale that ranged from 0 (definitely not) to 100 (definitely is) that each object in the pair was a blicket both before, midway, and after a trial for a total of three causal ratings. The exact points at which participants were asked to provide the three sets of ratings for objects A, B, and C is shown below in Table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Schematic of Experiment 2 | | | | | |
|  | **PRE** |  | **Mid** |  | **Post** |
| **Backwards-Blocking** | \* | AB++ | **\*** | A+ | **\*** |
| **Indirect Screening-Off** | \* | AB++ | **\*** | A- | **\*** |
| **One-Cause** | \* | A+;B- | **\*** | AB+ | **\*** |
| **Two-Cause** | \* | A+++ | **\*** | B-;B++ | **\*** |

Table 1. The three points at which participants were asked to provide pre-, mid-, and post-ratings in the across the four conditions. The asterisks correspond to where in a particular sequence participants were asked to provide causal ratings of objects A and B. The number of +/- signs corresponds to the number of times a particular sequence was demonstrated and whether the effect happened (+) or did not happen (-).

**Results**

All analyses were conducted in R (R Development Core Team, 2008). Figure 2 shows the mean pre- and post-causal ratings of objects A and B across the four conditions. Given evidence of non-normality and unequal variance in the causal-rating data, we used non-parametric analyses with 4,000 replications each for hypothesis testing and to estimate confidence intervals. In particular, the Shapiro-Wilks test indicated that the data were not statistically normally distributed, all *p*’s < .05. Likewise, the Levene’s test indicated heterogeneity of variance, *F*(3, 572) = 15.92, *p* < .001.

----------------------------------------------------------------

Insert Figure 2 about here

----------------------------------------------------------------

Because the principal goal of Experiment 1 was to examine to what extent adults engaged in BB and IS reasoning and to examine whether adults' causal ratings of the two objects confirmed the predictions of a Bayesian model, the traditional RW model, or the modified RW model, we only report the results for the BB and IS conditions. It is worth repeating that the 1C and 2C conditions serve strictly as control conditions to ensure that children (e.g., Sobel et al., 2004) and adults (e.g., Griffiths et al., 2011) understand the point of the experiment.

Note that causal likelihood ratings of both objects in each of the four conditions were also analyzed using linear mixed-effects models (LMM). This represents a better approach than either univariate ANOVA or ordinary least squares regression because it addresses unbalanced and non-independent designs and data (for an extended discussion, see Baayen, Davidson, & Bates, 2008).

**Indirect screening-off.** We first examined whether participants engaged in IS reasoning. A mixed-effects model was fit to participants' causal ratings in the IS condition. Participants were included as a random-effect factor and phase (pre vs. mid vs. post) and object type (A vs. B) were included as within-participants fixed-effects factors. This analysis yielded main effects for Phase, *F*(2,115) = 9.32, *p* < .001 and Object Type, *F*(1,115) = 260.49, *p* < .0001, which were qualified by a Phase x Object Type interaction, *F*(2,115) = 189.27, *p* < .0001.

Follow-up (permutation) planned comparisons for object A revealed that although participants' pre-ratings of object A (*M* = 47.58, Bootstrapped 95% CI[42.09, 53.07]) did not differ from their mid-ratings of object A (*M* = 56.13, Bootstrapped 95% CI[49.42, 62.83]), *p* = .82, Bootstrapped 95% CI[-17.26, 0.18]), participants provided higher pre-ratings of A than post-ratings of A (*M* = 1.88, Bootstrapped 95% CI[-0.38, 4.13]), *p* < .0001, Bootstrapped 95% CI[39.75, 51.66] and higher mid-ratings of A than post-ratings of it, *p* < .0001, Bootstrapped 95% CI[47.11, 61.39].

Planned comparisons for object B revealed that although participants provided lower pre-ratings of B (*M* = 51.33, Bootstrapped 95% CI[44.05, 58.62]) than post-ratings of B (*M* = 97.08, Bootstrapped 95% CI[94.14, 100.02]), *p* < .0001, Bootstrapped 95% CI[,-53.58, -37.92] and lower mid-ratings of B (*M* = 62.79, Bootstrapped 95% CI[55.48, 70.1] than post-ratings of it, *p* < .0001, Bootstrapped 95% CI[-42.12, -26.47], participants' pre- and mid-ratings of B did not differ.

**Backwards-blocking.** The final analysis examined whether adults engaged in BB reasoning. Similar to the preceding analysis, a mixed-effects model was fit to participants' causal ratings in the BB condition. Participants were included as a random-effect factor and phase (pre vs. mid vs. post) and object type (A vs. B) were included as within-participants fixed-effects factors. This analysis yielded main effects for Phase, *F*(2,115) = 17.21, *p* < .0001 and Object Type, *F*(1,115) = 21.61, *p* < .0001, which were qualified by a Phase x Object Type interaction, *F*(2,115) = 29.29, *p* < .0001.

Planned comparisons for object A revealed that although participants provided lower pre-ratings of A (*M* = 49.38, Bootstrapped 95% CI[42.33, 56.42]) than post-ratings of A (*M* = 98.54, Bootstrapped 95% CI[96.35, 100.73]) and lower mid-ratings of A (*M* = 57.5, Bootstrapped 95% CI[,47.83, 67.17]) than post-ratings of it, *p* < .0001, Bootstrapped 95% CI[-51.02, -31.06], participants' pre- and mid-ratings of A did not differ. A final set of planned comparisons for object B revealed that ratings of object B did not differ across the three rating phases, all p's > .2. Although adults’ ratings of objects A and B did not differ between at least two of the three rating phases, below we have included the predictions of each model to facilitate model comparison to determine with which model participants’ ratings of object B in the BB condition were consistent (see the Appendix for model derivation details).

**The simple Bayesian model predictions for BB.** Because Bayes’ rule is computed with respect to a space of competing causal models or hypotheses, it is necessary to show that space before presenting the predictions of the Bayesian model (see the Appendix for details about how this model was derived and for how the predictions were obtained). As such, Figure 3 below depicts the space of models for two objects (i.e., four possible hypotheses).

------------------------------------------------------------------------------------------------------------

Insert Figure 3 about here

------------------------------------------------------------------------------------------------------------

The predictions corresponding to the above hypothesis space are shown below in Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 1 BAYESIAN MODEL PREDICTIONS** | | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ (*p*(*h|d*))** | **After A+ (*p*(*h|d*))** |
| Model 0 | *p*2 | *p*/(2-*p*) | *P* |
| Model 1 | *p*(1-*p*) | (1-*p*)/(2-*p*) | 1-*p* |
| Model 2 | *p*(1-*p*) | (1-*p*)/(2-*p*) | 0 |
| Model 3 | (1-*p*)2 | 0 | 0 |
| **Probabilities of objects A and B** | | | |
| Object A | *p* | 1/(2-*p*) | 1 |
| Object B | *p* | 1/(2-*p*) | *p* |

Table 2. This table displays the predictions of the Bayesian model for how adults should rate objects A and B in Experiment 1 before the BB, after the AB+ event, and then after the A+ event. In addition, this table displays the posterior probabilities assigned to each graph before the BB event, after the AB+ event, and then after the A+ event. The prior probabilities of objects A and B—which represent predictions for how adults should rate both objects—were derived according to the principle of object independence. The subsequent posterior probabilities are computed by dividing the prior for that graph by the sum of the prior probabilities for the graphs in which a causal link exists for that particular object.

**The traditional RW model predictions for BB.** The predictions of the traditional RW model applied to the BB condition are shown below in Table 3.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 1 RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 1 |
| B | *p* | *p* | *p* |

Table 3. This table displays the traditional Rescorla-Wagner model’s predictions of the causal ratings for objects A and B before the BB event, after the AB+ event, and then after the A+ event. The ratings in each cell can be thought to represent the confidence that a particular object is a cause.

**The modified RW model predictions for BB.** The predictions of the modified RW model applied to the BB condition are shown below in Table 4.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. **EXPERIMENT 1 MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 1 |
| B | *p* | *p* | 0 |

Table 4. This table displays the modified Rescorla-Wagner model’s predictions of the causal ratings for objects A and B before the BB event, after the AB+ event, and then after the A+ event. The ratings in each cell can be thought to represent the confidence that a particular object is a cause.

It should be clear from these predictions that adults’ causal ratings were not consistent with the predictions of the Bayesian model. This is because participants’ pre- and mid-ratings of objects A and B did not differ. In contrast, the simple Bayesian model predicts that participants’ mid-ratings of objects A and B will be higher than their pre-ratings of it. This suggests that a Bayesian-inference mechanism may not underpin adults’ processing of the BB task in the context of two objects. To determine whether this conclusion was warranted—that is, whether the absence of any differences between the pre-, mid-, and post-ratings of objects A and B were statistically meaningful—we used the R package “BayesFactor” to compute a two-sided, paired-samples Bayes’ factor, where the prior was assumed to follow a Cauchy distribution with a scale parameter, λ, set to 0.7. For reference, the scale parameter corresponds roughly to a medium effect size.

This analysis revealed that the observed difference between participants’ pre- and mid-ratings of object A was 3.2 times more likely under the alternative hypothesis (i.e., that the difference is reliable) than under the null hypothesis (i.e., that there is no reliable difference). Likewise, the observed difference between participants’ pre- and mid-ratings of object B was 0.42 times more likely under the null hypothesis than under the alternative hypothesis. Furthermore, the observed difference between participants’ pre- and post-ratings was 0.31 times more likely under the null hypothesis than under the alternative hypothesis. Finally, the observed difference between participants’ mid- and post-ratings of object B was 0.45 times more likely under the null hypothesis than under the alternative hypothesis. With the exception of the observed difference between the pre- and mid-ratings of object A—which represents weak evidence in favor of the alternative hypothesis—Bayes’ factors between 0.3 and 0.45 represent substantial support in favor of the null hypothesis that the observed differences were not reliably different. Bayes’ factors are particularly useful because, unlike frequentist *p*-values, they enable one to accept or reject the null hypothesis and to infer whether the data are insensitive perhaps because of insufficient statistical power (for discussions see Jeffreys, 1961, and Raftery, 1995).

The modified RW model also does not provide the best fit to adults’ causal ratings of objects A and B in the BB condition given that this model does not account for adults’ treatment of object B across all three rating phases. Indeed, whereas the modified RW model predicts that participants’ causal ratings of object B should drop between the pre- and post-rating phases, participants’ actual ratings remained constant across all three rating phases. However, participants’ ratings of objects A and B were consistent with the predictions of the traditional RW model. This is evidenced by the fact that, in the BB condition, adults’ causal ratings of objects A and B were consistent with the predictions of the traditional RW model across all three rating phases. As such, the traditional RW model, but neither the simple Bayesian model nor the modified RW model, provides the best fit to the present BB causal-rating data.

**Evidence of a Bayesian inference mechanism**

Given that adults' causal ratings of object B did not change across the three rating phases of the BB condition, the results presented thus far suggest that adults do not engage in BB reasoning in a manner that is consistent with the predictions of a simple Bayesian model. However, before it can be concluded definitively that adults do not rely on a Bayesian mechanism to reason about causal events in the context of two objects, it is necessary to show that their causal ratings of objects A and B also do not conform with the predictions of the Bayesian model for the IS condition across the three rating phases. Thus, similar to above, we present the predictions of the three models.

**The simple Bayesian model predictions for IS.** The predictions of the modified RW model applied to the BB condition are shown below in Table 5.

|  |  |  |  |
| --- | --- | --- | --- |
| **BAYESIAN MODEL PREDICTIONS: IS CONDITION** | | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ (*p*(*h|d*))** | **After A+ (*p*(*h|d*))** |
| Model 0 | *p*2 | *p*/(2-*p*) | 0 |
| Model 1 | *p*(1-*p*) | (1-*p*)/(2-*p*) | 0 |
| Model 2 | *p*(1-*p*) | (1-*p*)/(2-*p*) | 1 |
| Model 3 | (1-*p*)2 | 0 | 0 |
| **Probabilities of objects A and B** | | | |
| Object A | *p* | 1/(2-*p*) | 0 |
| Object B | *p* | 1/(2-*p*) | 1 |

Table 5. This table displays the predictions of the Bayesian model for how adults should rate objects A and B before the IS event, after the AB+ event, and then after the A- event. In addition, this table displays the posterior probabilities assigned to each graph before the IS event, after the AB+ event, and then after the A- event. The prior probabilities of the four models were derived according to the principle of object independence. The subsequent posterior probabilities are computed by dividing the prior for that graph by the sum of the prior probabilities for the graphs in which a causal link exists for that particular object.

**The traditional RW model predictions for IS.** These predictions of the traditional RW model for the IS condition can be seen below in Table 6.

|  |  |  |  |
| --- | --- | --- | --- |
| **RESCORLA-WAGNER MODEL PREDICTIONS: IS CONDITION** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 0 |
| B | *p* | *p* | *p* |

Table 6. This table displays the predictions of the traditional RW model for objects A and B in the IS condition across the three rating phases.

**The modified RW model predictions for IS.** The predictions of the modified RW model are shown below in Table 7.

|  |  |  |  |
| --- | --- | --- | --- |
| **MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS: IS CONDITION** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 0 |
| B | *p* | *p* | 1 |

Table 7. This table displays the predictions of the modified RW model for objects A and B in the IS condition across all three rating phases.

As can be seen from Tables 5-7 above, participants’ ratings of both objects did confirm with the predictions of the modified RW model but not with the predictions of the simple Bayesian model or the traditional RW model. There are two reasons for this. First, participants’ pre- and mid-ratings of objects A and B did not differ from each other in both the BB and IS conditions. Second, participants’ ratings of object B increased between the mid- and post-rating phase in the IS condition but remained unchanged across all three rating phases in the BB condition. The distribution of responses to object A also accorded with the predictions of the modified RW model, but not with the predictions of the other two models, across all three rating phases in both the IS and BB conditions. These results suggest that, together, the traditional RW model and modified RW model provided a better fit to the data than did the Bayesian model, although neither model alone accounted for adults’ causal-rating performance in both conditions. Nonetheless, the results from Experiment 1 suggest that an associative mechanism, rather than a Bayesian mechanism, may underlie adult causal reasoning for two objects.

Discussion

The aim of Experiment 1 was to test the predictions of the traditional RW model, the modified RW model, and a Bayesian model for causal reasoning with two objects. The results revealed that adults' responses in the BB and IS conditions were consistent with the predictions of the modified and traditional RW models but not with the Bayesian model. Taken together, these results indicate that adults did not engage in BB reasoning; that is, they do not block object B such that B undergoes a drop between any two of the three rating phases as would be predicted either by the modified RW model or a simple Bayesian model. That adults’ ratings did not conform with the predictions of the simple Bayesian model further suggests that a Bayesian-inference mechanism may not underpin adult causal reasoning in a blicket-detector context that consists of two objects. It will be recalled that if such a mechanism did underpin adult causal reasoning, then their rating of object B should have increased between the pre- and mid-rating phases but then return to baseline (i.e., the pre-rating value) between the mid- and post-rating phases. Likewise, if this mechanism underpinned adults' processing of the IS event, then their ratings of objects A and B should have increased between the pre- and mid-rating phase. However, their ratings of object A should have dropped between the mid- and post-rating phases, whereas their ratings of object B should have increased between these two phases. One plausible objection to this conclusion is that the reason the adults did not engage in BB reasoning was simply because the experiment was underpowered. However, this criticism is weakened by the fact that an a priori power analysis indicated that 16 participants would be sufficient to have an 80% chance of detecting a medium-size effect assuming that such an effect existed. This conclusion is also weakened by the fact that a Bayes’ factor analysis indicated that the observed differences between participants’ pre-, mid-, and post-ratings of objects A and B represented substantial evidence in support for the null than under the alternative hypothesis.

This is the first series of experiments to use the blicket-detector task to show that adults do not engage in BB reasoning and to show that adults use associative processes more generally to reason about causal events that consist of two objects in a blicket-detector context. Although the present data suggests that adults use associative processes to reason about causal events, an open question concerns why their responses were in line with the predictions of the traditional RW model in the BB condition but with the predictions of the modified RW model in the IS condition. One speculative possibility is that adults relied on propositional reasoning to assess the present causal events (for discussions see De Houwer, Beckers, & Vandorpe, 2005; Lovibond, 2003; Mitchell, De Houwer, & Lovibond, 2009; Mitchell, Killedar, & Lovibond, 2005). On this account, upon seeing that object A failed to make the machine go by itself in the IS condition (after observing the initial AB+ phase), the adults may have concluded that object B must be a blicket. In contrast, upon observing that object A made the machine activate in the BB condition, the adults may have concluded that B may or may not be a cause. These two explanations—perhaps in conjunction with a domain-general associative-learning mechanism—can explain why adults’ causal ratings were consistent with the predictions of the traditional RW model in the BB condition but with the modified RW model in the IS condition. We return to this issue in the General Discussion. Nonetheless, the fact that adults’ responses were inconsistent with the predictions of a Bayesian model across the BB and IS conditions suggests at a minimum that a Bayesian-inference may not underpin adult causal reasoning about two objects in a blicket-detector context.

An unresolved question from Experiment 1 concerns to what extent adults will engage in BB reasoning—in a manner either consistent with the predictions of a simple Bayesian model or the modified RW model—when asked to reason about more than two objects. Although Experiment 1 demonstrated that adults do not engage in BB reasoning when asked to reason about two objects in the blicket-detector context, it is unclear whether adults will engage in BB reasoning when reasoning about more than two objects. Perhaps more importantly, it is also unclear whether Sobel and Kirkham’s (2006) BB finding that participants blocked a redundant cause and considered a third object—which was never demonstrated on the machine—to be more of a cause of the machine’s activation than the redundant cause would generalize to the present experiment that tested adults and used three objects. These are important concerns to address because successful causal reasoning in the real world requires reasoning about multiple competing causes and because it is unknown what role, if any, non-acted-upon objects have on acted-on objects. This latter issue is also important to address because it is possible further to determine which of the three models considered here provides the best account of adults’ causal performance. This is because each model makes a different prediction for the three objects.

**Experiment 2**

The goal of Experiment 2 was to determine whether adults’ causal ratings of three distinct objects in the IS and BB conditions are consistent with the predictions of the three models.

**Method**

**Participants.** Twenty-four college students were tested in Experiment 1 (13 females, 11 males). These participants received credit for an introductory-level psychology course.

**Stimuli, Design, and Procedure.**  All aspects of Experiment 2 were identical to Experiment 1 with one exception: Twelve objects were used to demonstrate the effect, although participants only ever experienced three objects. These included cube, cylinder, and triangle objects, each differently colored and approximately 1" in diameter. Within each trial, the objects were different shapes and colors and the object that was designated as the blicket was counterbalanced across participants. Further, the procedure used in Experiment 2 was identical to Experiment 1 except that, across all four trials, object C never participated on the machine and remained on the table in plain sight throughout an entire trial. The exact points at which participants were asked to provide the three sets of ratings for objects A, B, and C is identical to those in Table 6.

**Results**

Given evidence of non-normality and unequal variance in the causal-rating data, all analyses used non-parametric analyses with 4,000 replications each for hypothesis testing and to estimate confidence intervals. In particular, the Shapiro-Wilks test indicated that the data were not statistically normally distributed, all *p*’s < .05. Likewise, the Levene’s test indicated heterogeneity of variance, *F*(3, 860) = 6.28, *p* < .001. Figure 4 shows the mean pre and post causal ratings of objects A, B, and C across the four conditions.

----------------------------------------------------------------

Insert Figure 4 about here

----------------------------------------------------------------

**Indirect screening-off.** The first analysis examined whether participants engaged in IS reasoning. A mixed-effects model was fit to participants' causal ratings in the IS condition. Participants were included as a random-effects factor and Phase (pre vs. mid vs. post) and Object Type (A vs. B vs. C) were included as within-participants fixed-effects factors. This analysis yielded main effects for Phase, *F*(2,184) = 6.13, *p* < .005 and Object Type, *F*(2,184) = 49.9, *p* < .0001, which was qualified by a Phase x Object Type interaction, *F*(4,184) = 48.21, *p* < .0001.

Follow-up (permutation) planned comparisons for object A revealed that participants provided higher pre-ratings (*M* = 47.58, Bootstrapped 95% CI[42.09, 53.07]) and mid-ratings of object A (*M* = 65.83, Bootstrapped 95% CI[57.94, 73.73]) than post-ratings of object A (*M* = 10.21, Bootstrapped 95% CI[2.23, 18.19]), both *p*'s < .0001. Likewise, participants provided higher mid-ratings of object A than pre-ratings of it, Bootstrapped 95% CI[-21.11, -1.39].

Planned comparisons for object B revealed that participants provided lower pre- (*M* = 52.08, Bootstrapped 95% CI[44.26, 59.91]) and mid-ratings of object B (*M* = 67.92, Bootstrapped 95% CI[59.40, 76.43]) than post-ratings of it (*M* = 98.75, Bootstrapped 95% CI[97.55, 99.95]), both p's < .0001. Participants also provided lower pre-ratings of object A than mid-ratings of it, p < .0001, Bootstrapped 95% CI[-27.54, -4.13]. In contrast to participants' causal ratings of objects A and B, a final set of planned comparisons for object C revealed that participants' ratings of object C did not differ across the three rating phases, all *p*'s > .61.

**Backwards-blocking.** The second analysis examined whether participants engaged in BB reasoning. A mixed-effects model was fit to participants' causal ratings in the BB condition. Similar to the previous analysis, participants were included as a random-effects factor and Phase (pre vs. mid vs. post) and Object Type (A vs. B vs. C) were included as within-participants fixed-effects factors. This analysis yielded main effects for Phase, *F*(2,184) = 8.73, *p* < .001 and Object Type *F*(2,184) = 22.25, *p* < .0001, which was qualified by a Phase x Object Type interaction, *F*(4,184) = 11.96, *p* < .0001.

Follow-up (permutation) planned comparisons for object A revealed that participants provided higher post-ratings of object A (*M* = 92.96, Bootstrapped 95% CI[84.18, 101.73]) than either mid-ratings (*M* = 60.67, Bootstrapped 95% CI[51.61, 69.72]) or pre-ratings of it (*M* = 47.96, Bootstrapped 95% CI[41.68, 54.24]), both *p*'s < .0001. Similarly, participants provided higher mid-ratings of object A than pre-ratings of it, *p* < .0001.

Planned comparisons for object B revealed that although participants provided higher mid-ratings of object B (*M* = 57.71, Bootstrapped 95% CI[47.99, 67.42]) than pre-ratings (*M* = 45.42, Bootstrapped 95% CI[40.49, 50.34]) of it, *p* < .0001, Bootstrapped 95% CI[-23.18, -1.4], participants' pre- and post-ratings (*M* = 45.21, Bootstrapped 95% CI[34.08, 56.34]) of B did not differ reliably, *p* = .97, Bootstrapped 95% CI[-11.91, 12.33]. In contrast, there was a marginal difference between participants' mid- and post-ratings of B, *p* = .15, Bootstrapped 95% CI[-2.06, 27.07].

To determine whether this marginal difference was practically meaningful, we conducted a Bayes’ factor analysis in the same manner as that in Experiment 1. This analysis indicated that the observed difference between the mid- and post-ratings of B (Δ12.5) were 3.89 times more likely under the alternative hypothesis than under the null hypothesis. This represents positive evidence in support of the alternative hypothesis that participants provided reliably higher mid-ratings of B than post-ratings of it. A final set of comparisons for object C revealed that participants' ratings of object C did not differ across the three rating phases, all *p*’s > .75.

Similar to Experiment 1, we present the predictions of each of the three models below to determine with which model’s predictions participants’ ratings of objects A-C accorded.

**The simple Bayesian model predictions for BB.** The sample space—which is comprised of eight hypotheses for all three objects—is shown below in Figure 4 and the predictions of this model are shown below in Table 8.

----------------------------------------------------------------

Insert Figure 5 about here

----------------------------------------------------------------

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 2 BAYESIAN MODEL PREDICTIONS (3 Objects)** | | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ (*p*(*h|d*))** | **After A+ (*p*(*h|d*))** |
| Model 0 | *p*3 | *p*2/(2-*p*) | *p*2 |
| Model 1 | *p*2(1-*p)* | *p*(1-*p*)/(2-*p*) | *p*(1-*p*) |
| Model 2 | *p*2(1-*p)* | *p*(1-*p*)/(2-*p*) | *p*(1-*p*) |
| Model 3 | *p*(1-*p*)2 | (1-*p*)2/(2-*p*) | (1-*p*)2 |
| Model 4 | *p*2(1-*p*) | *p*(1-*p*)/(2-*p*) | 0 |
| Model 5 | *p*(1-*p*)2 | (1-*p*)2/(2-*p*) | 0 |
| Model 6 | *p*(1-*p*)2 | 0 | 0 |
| Model 7 | (1-*p*)3 | 0 | 0 |
| **Probabilities of objects A, B, C, and D** | | | |
| Object A | *p* | 1/(2-*p*) | 1 |
| Object B | *p* | 1/(2-*p*) | *p* |
| Object C | *p* | *p* | *p* |

Table 8. This table displays the predictions of the Bayesian model for how adults should rate objects A, B, and C in Experiment 2 before the BB event, after the AB+ event, and then after the A+ event. In addition, this table displays the posterior probabilities assigned to each graph before the BB event, after the AB+ event, and then after the A+ event. The prior probabilities of objects A, B, and C—which represent predictions for how adults should rate both objects—were derived according to the principle of object independence and can be found in the Appendix. The subsequent posterior probabilities are computed by dividing the prior for that graph by the sum of the prior probabilities for the graphs in which a causal link exists for that particular object.

**The traditional RW model for BB.** The predictions of the traditional RW model applied to the BB condition for three objects are shown below in Table 9.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 2 RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | *p* |
| C | *p* | *p* | *p* |

Table 9. This table displays the predictions of the traditional RW model for objects A, B, and C in the BB condition across all three rating phases.

**The modified RW model for BB.** The predictions of the modified RW model applied to the BB condition for three objects are shown below in Table 10.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 2 MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | 0 |
| C | *p* | *p* | *p* |

Table 10. This table displays the predictions of the modified RW model for objects A, B, and C in the BB condition across all three rating phases.

It should be clear from Tables 8 to 10 that although participants’ pre- and post-ratings of object A and pre-, mid-, and post-ratings of object C were consistent with the predictions of the modified RW and traditional RW models, participants’ ratings of object A and B did not accord with predictions of both models at other points. This indicates that neither the traditional RW model nor the modified RW model provides the best account of how adults reason about the BB condition in the context of three objects. In contrast, it is clear that the simple Bayesian model provides the best account of adults’ causal ratings of objects A-C. This is based on the fact that their causal ratings of objects A-C were consistent with the predictions of the model across all three rating phases.

**Evidence of a Bayesian inference mechanism**

The final analysis examined whether participants used a Bayesian mechanism to reason about the IS condition. Thus, the predictions of the Bayesian model applied to the IS condition for three objects are shown below in Table 11.

|  |  |  |  |
| --- | --- | --- | --- |
| **BAYESIAN MODEL PREDICTIONS: IS CONDITION (3 Objects)** | | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ (*p*(*h|d*))** | **After A+ (*p*(*h|d*))** |
| Model 0 | *p*3 | *p*2/(2-*p*) | 0 |
| Model 1 | *p*2(1-*p)* | *p*(1-*p*)/(2-*p*) | 0 |
| Model 2 | *p*2(1-*p)* | *p*(1-*p*)/(2-*p*) | 0 |
| Model 3 | *p*(1-*p*)2 | (1-*p*)2/(2-*p*) | 0 |
| Model 4 | *p*2(1-*p*) | *p*(1-*p*)/(2-*p*) | *p* |
| Model 5 | *p*(1-*p*)2 | (1-*p*)2/(2-*p*) | 1-*p* |
| Model 6 | *p*(1-*p*)2 | 0 | 0 |
| Model 7 | (1-*p*)3 | 0 | 0 |
| **Probabilities of objects A, B, C, and D** | | | |
| Object A | *p* | 1/(2-*p*) | 0 |
| Object B | *p* | 1/(2-*p*) | 1 |
| Object C | *p* | *p* | *p* |

Table 11. This table displays the predictions of the Bayesian model applied to the IS conditions for objects A, B, and C. This model was derived in the same was at that for the BB condition.

The predictions of the traditional RW model for objects A, B, and C are shown below in Table 12.

|  |  |  |  |
| --- | --- | --- | --- |
| **RESCORLA-WAGNER MODEL PREDICTIONS: IS CONDITION** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 0 |
| B | *p* | *p* | *p* |
| C | *p* | *p* | *p* |

Table 12. This table displays the predictions of the traditional RW model for objects A, B, and C in the IS condition across the three rating phases.

The predictions of the modified RW model for objects A, B, and C are shown below in Table 13.

|  |  |  |  |
| --- | --- | --- | --- |
| **MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS: IS CONDITION** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 0 |
| B | *p* | *p* | 1 |
| C | *p* | *p* | *p* |

Table 13. This table displays the predictions of the modified RW model for objects A, B, and C in the IS condition.

In light of the predictions of the three models in Tables 11-13, it is clear that although adults’ causal ratings of all three objects neither conformed with the predictions of the traditional RW model nor with the predictions of the modified RW model, they did conform with the predictions of the simple Bayesian model. This is based on the fact that participants’ ratings of objects A and B increased between the pre- and mid-rating phases across the BB and IS conditions—which is a unique prediction of the simple Bayesian model—and changed in a manner that was consistent with the same model between the mid- and post-rating phases for objects A and B. Moreover, that participants’ causal ratings of objects A, B, and C were consistent with the predictions of the Bayesian model across all three rating phases in both the BB and IS conditions suggests that adults may have used Bayesian inference to reason about the present events.

**Discussion**

The results of Experiment 2 confirmed the predictions of the Bayesian model for both the BB and IS conditions. These findings are important for two reasons. First, the fact that adults’ causal ratings conformed with the predictions of the Bayesian model but not with the predictions of either of the two associative models suggests, albeit tentatively, that a Bayesian mechanism may underpin human causal reasoning about three, but not two, objects. Second, these findings suggest that BB reasoning may be best understood as a return to baseline in the causal rating of object B between a mid- and post-rating phase rather than an absolute drop in the rating of object B between a pre- and post-rating phase for three objects.

To our knowledge, this is the first study systematically to examine to what extent adults’ causal ratings of three objects are consistent with the predictions of one of three competing models and to show that a simple Bayesian model provides the best account of adults’ causal ratings of three objects in both the BB and IS conditions. This is also the first study to show that participants are not more likely to consider a third, non-acted upon object to be more of a cause than a redundant cause. This is based on the fact that participants’ ratings of object C—which was never demonstrated on the machine—did not change across the three rating phases in either the BB or IS conditions.

These results notwithstanding, key questions concern (1) why adults processed the events associatively in Experiment 1 but used Bayesian inference to reason about the events in Experiment 2 and (2) whether adults would continue to provide causal ratings that are consistent with the predictions of a Bayesian model when they are asked to reason about more than three objects. Experiment 3 was designed to address this second issue and the first issue is addressed in detail in the General Discussion. Note that the second issue is worth investigating principally because causal reasoning in the real world involves reasoning about many different competing causes. In this way, one can determine the viability of a Bayesian-inference mechanism by assessing to what extent it can account for how adults reason about more than two or three competing causes.

**Experiment 3**

The aims of Experiment 3 were threefold. The first aim was to replicate the observed drop in the rating of B between the mid- and post-rating phases in the BB condition in Experiment 2 to determine its reliability and to examine in what ways adults reason about four objects. The second aim was to determine to what extent adults' causal ratings of four objects are consistent with the predictions of a Bayesian model or one of two associative models. The third aim was to determine whether a third and fourth object—which, importantly, would not participate on the machine—impacts adults’ assessments of an acted-on first or second object.

**Method**

**Participants, stimuli, and design.** Twenty students were tested in Experiment 2 (13 males, 7 females). The device used in Experiment 3 was the same as that used in Experiments 1 and 2. However, sixteen rather than either or twelve objects were used to demonstrate the effect, although a given participant only experienced four objects. These included cubes, cylinders, triangular pyramids, and rectangular prisms objects, each differently colored and approximately 1" in diameter. Similar to Experiment 1, the object that was designated as the blicket was counterbalanced across participants to ensure that the effect could not be attributed to a particular object.

**Procedure.** The procedure, pretest phase, four test trials, and rating scale were identical to those used in Experiments 1 and 2 with one key exception. Rather than rating the likelihood that each of three objects were blickets on a scale that ranged between 0 (definitely not) to 100 (definitely) as was done in Experiment 2, participants were asked to rate the likelihood that each of four objects—namely, objects A, B, C, and D—were blickets both before a particular test trial (e.g., backwards blocking, IS, 1C, and 2C), at a midway point, and then after a particular trial was demonstrated. Note that although objects A and B were used to demonstrate each test event, objects C and D remained on the table throughout each test trial and never participated on the machine.

**Results**

Adults’ mean causal ratings in all four conditions across the three rating phases for objects A, B, C, and D are presented in Figure 5. Sex or condition did not interact with any of the factors, and therefore we collapsed across these factors. Likewise, because of evidence of nonnormality, *p < .*05, and heteroskedasticity, *p < .*01 based on the Shapiro-Wilks test and the Levene’s test, all analyses reported in Experiment 4 used non-parametric analyses with 4,000 replications each for hypothesis testing and to estimate confidence intervals.

----------------------------------------------------------------

Insert Figure 6 about here

----------------------------------------------------------------

**Indirect screening-off.** The first analysis examined whether adults engaged in IS reasoning. A mixed-effects model was fit to participants' causal ratings in the IS condition. Participants were included as a random-effects factor and Phase (pre vs. mid vs. post) and Object Type (A vs. B vs. C) were included as within-participants fixed-effects factors. This analysis yielded main effects for Phase, *F*(2,209) = 6.46, *p* < .005 and Object Type, *F*(2,209) = 28.48, *p* < .0001, which were qualified by a Phase x Object Type interaction, *F*(6,209) = 40.93, *p* < .0001.

Follow-up (permutation) planned comparisons for object A revealed that participants provided lower post-ratings of object A (*M* = 8.25, Bootstrapped 95% CI[-1.86,18.36]) than either mid- (*M* = 78.15, Bootstrapped 95% CI[71.46,84.84]) or pre-ratings of it (*M* = 53.5, Bootstrapped 95% CI[48.8,62.19]), both *p*'s < .0001. Furthermore, participants provided higher mid-ratings of object A than pre-ratings of it, *p* < .0001, Bootstrapped 95% CI[-35.59,-13.71].

Planned (permutation) comparisons for object B revealed that participants provided higher post-ratings of object B (*M* = 99.5, Bootstrapped 95% CI[98.53,100.47]) than either mid- (*M* = 64.4, Bootstrapped 95% CI[51.92,76.88]) or pre-ratings of it (*M* = 51.75, Bootstrapped 95% CI[42.89,60.61]), both *p*'s < .0001. Likewise, participants mid-ratings differed reliably from their pre-ratings of it, *p* < .0001, Bootstrapped 95% CI[-27.76,2.47].

A final set of analyses was conducted to determine whether participants' ratings of objects C and D differed across the three rating phases. This analysis indicated that neither did participants' ratings of object C across all three rating phases differ reliably nor did their ratings of object D, all *p*'s > .38.

**Backwards-blocking.** The final analysis examined whether adults engaged in BB reasoning. Similar to the preceding analysis, a mixed-effects model was fit to participants' causal ratings in the BB condition. Participants were included as a random-effect factor and phase (pre vs. mid vs. post) and object type (A vs. B) were included as within-participants fixed-effects factors. This analysis yielded main effects for Phase, *F*(2,209) = 11.38, *p* < .0001 and Object Type, *F*(3,209) = 31.76, *p* < .0001, which were qualified by a Phase x Object Type interaction, *F*(6,209) = 16.09, *p* < .0001.

Planned (permutation) planned comparisons for object A revealed that participants provided higher post-ratings of object A (*M* = 99.75, Bootstrapped 95% CI[99.25,100.25]) than either mid-ratings (*M* = 71, Bootstrapped 95% CI[64.65,77.35]) or pre-ratings (*M* = 50.25, Bootstrapped 95% CI[45.26,55.24]) of it, both *p*'s < .0001. Similarly, participants provided reliably higher mid-ratings of object A than pre-ratings of it, *p* < .0001, Bootstrapped 95% CI[-28.73,-12.77]).

Planned (permutation) planned comparisons for object B revealed that although participants' pre- (*M* = 50.9, Bootstrapped 95% CI[44.56,57.24]) and post-ratings (*M* = 50.25, Bootstrapped 95% CI[37.71,62.79]) did not differ reliably, *p* = .94, Bootstrapped 95% CI[-13.36,14.66], participants provided reliably higher mid-ratings of object B (*M* = 63.4, Bootstrapped 95% CI[54.73,72.07]) than pre-ratings of it, *p* < .0001, Bootstrapped 95% CI[-23.27,-1.73]. Finally, participants' mid- and post-ratings of object B differed only marginally, *p* = .1, Bootstrapped 95% CI[-2.16,28.46]. To determine the meaningfulness of this difference, we estimated a Bayes' factor in the same manner as that in Experiments 1 and 2. This analysis indicated that the data were 3.24 times more likely under the alternative hypothesis—that participants' mid- and post-ratings of object B differed reliably—than under the null hypothesis—that these ratings did not differ reliably. This represents substantial evidence that the two ratings are meaningfully different.

The next analysis was undertaken to determine whether participants' ratings of objects C and D differed across the three rating phases. This analysis indicated that neither did participants' ratings of object C across all three rating phases differ reliably nor did their ratings of object D, all *p*'s > .48. These ratings are consistent with the predictions of the simple Bayesian model applied to the BB condition but not with the predictions of either of the two associative models applied to the same condition.

In a similar vein to Experiments 1 and 2, below we present the predictions of each of the three models to ascertain with which of the three computational and analytical models participants’ treatment of all four objects accorded.

**The simple Bayesian model predictions for BB.** The sample space—which is comprised of eight hypotheses for all three objects—is shown below in Figure 4 and the predictions of this model are shown below in Table 14.

----------------------------------------------------------------

Insert Figure 7 about here

----------------------------------------------------------------

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 3 BAYESIAN MODEL PREDICTIONS (4 Objects)** | | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ (*p*(*h|d*))** | **After A+ (*p*(*h|d*))** |
| Model 0 | *p*4 | *p*3/(2-*p*) | *p*3 |
| Model 1 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | *p*2(1-*p*) |
| Model 2 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | *p*2(1-*p*) |
| Model 3 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | *p*(1-*p*)2 |
| Model 4 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | *p*2(1-*p*) |
| Model 5 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | *p*(1-*p*)2 |
| Model 6 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | *p*(1-*p*)2 |
| Model 7 | *p*(1-*p*)3 | (1-p)3/(2-*p*) | (1-*p*)3 |
| Model 8 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | 0 |
| Model 9 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | 0 |
| Model 10 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | 0 |
| Model 11 | *p*(1-*p*)3 | (1-p)3/(2-*p*) | 0 |
| Model 12 | *p*2(1-*p*)2 | 0 | 0 |
| Model 13 | *p*(1-*p*)3 | 0 | 0 |
| Model 14 | *p*(1-*p*)3 | 0 | 0 |
| Model 15 | (1-*p*)4 | 0 | 0 |
| **Probabilities of objects A, B, C, and D** | | | |
| Object A | *p* | 1/(2-*p*) | 1 |
| Object B | *p* | 1/(2-*p*) | *p* |
| Object C | *p* | *p* | *p* |
| Object D | *p* | *p* | *p* |

Table 14. This table displays the predictions of the Bayesian model for how adults should rate objects A, B, C, and D in Experiment 3 before the BB, after the AB+ event, and then after the A+ event. In addition, this table displays the posterior probabilities assigned to each graph before the BB event, after the AB+ event, and then after the A+ event. The prior probabilities of objects A, B, C, and D—which represent predictions for how adults should rate both objects—were derived according to the principle of object independence. The subsequent posterior probabilities are computed by dividing the prior for that graph by the sum of the prior probabilities for the graphs in which a causal link exists for that particular object.

Table 15. This table displays the predictions of the traditional RW model for objects A, B, C, and D in the BB condition in Experiment 3.

**The traditional RW model for BB.** The predictions of the traditional RW model applied to the BB condition for four objects are shown below in Table 15.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 3 RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | *p* |
| C | *p* | *p* | *p* |
| D | *p* | *p* | *p* |

**The modified RW model for BB.** predictions of the modified RW model applied to the BB condition for four objects are shown below in Table 16.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 3 MODIFIED RESCORLA-WAGNER MODEL PREDICITONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | 0 |
| C | *p* | *p* | *p* |
| D | *p* | *p* | *p* |

Table 16. This table displays the predictions of the modified RW model for objects A, B, C, and D in the BB condition in Experiment 3.

In light of these predictions, it should be clear that the simple Bayesian model provided the best fit to the data in Experiment 3. This is based on the fact that adults’ causal ratings of objects A-D were consistent with the predictions of the simple Bayesian model across all three rating phases but neither with the predictions of the traditional RW model nor with those of the modified RW models. This suggests that adults engaged in BB reasoning in a manner that was consistent with the predictions of the simple Bayesian model.   
**Evidence of a Bayesian inference mechanism**

Similar to the preceding two experiments, the final analysis examined to what extent adults’ causal ratings were consistent with the predictions of the simple Bayesian model or with either of the two associative models applied to the IS condition. The predictions of the Bayesian model applied to the IS condition for four objects are shown below in Table 17.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 2 BAYESIAN MODEL PREDICTIONS (4 Objects)** | | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ (*p*(*h|d*))** | **After A+ (*p*(*h|d*))** |
| Model 0 | *p*4 | *p*3/(2-*p*) | 0 |
| Model 1 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | 0 |
| Model 2 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | 0 |
| Model 3 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | 0 |
| Model 4 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | 0 |
| Model 5 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | 0 |
| Model 6 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | 0 |
| Model 7 | *p*(1-*p*)3 | (1-p)3/(2-*p*) | 0 |
| Model 8 | *p*3(1-*p*) | *p*2(1-*p*)/(2-*p*) | *P*2 |
| Model 9 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | *p(1-p)* |
| Model 10 | *p*2(1-*p*)2 | *p*(1-*p*)2/(2-*p*) | *p(1-p)* |
| Model 11 | *p*(1-*p*)3 | (1-p)3/(2-*p*) | (1-*p*)2 |
| Model 12 | *p*2(1-*p*)2 | 0 | 0 |
| Model 13 | *p*(1-*p*)3 | 0 | 0 |
| Model 14 | *p*(1-*p*)3 | 0 | 0 |
| Model 15 | (1-*p*)4 | 0 | 0 |
| **Probabilities of objects A, B, C, and D** | | | |
| Object A | *p* | 1/(2-*p*) | 0 |
| Object B | *p* | 1/(2-*p*) | 1 |
| Object C | *p* | *p* | *p* |
| Object D | *p* | *p* | *p* |

Table 17. This table displays the predictions of the Bayesian model applied to the IS conditions for objects A, B, C, and D. This model was derived in the same was at that for the BB condition.

The predictions of the traditional RW model for objects A, B, C, and D are shown below in Table 18.

|  |  |  |  |
| --- | --- | --- | --- |
| **RESCORLA-WAGNER MODEL PREDICTIONS: IS CONDITION** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 0 |
| B | *p* | *p* | *p* |
| C | *p* | *p* | *p* |
| D | *p* | *p* | *p* |

Table 18. This table displays the predictions of the modified RW model for objects A, B, C, and D in the IS condition across all three rating phases.

The predictions of the modified RW model for objects A, B, C, and D are shown below in Table 19.

|  |  |  |  |
| --- | --- | --- | --- |
| **MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS: IS CONDITION** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 0 |
| B | *p* | *p* | 1 |
| C | *p* | *p* | *p* |
| D | *p* | *p* | *p* |

Table 19. This table displays the predictions of the modified RW model for objects A, B, C, and D in the IS condition.

Based on these predictions presented in Tables 17-19, it is evident that adults’ causal ratings of all four objects conformed with the predictions of the simple Bayesian model applied to the IS condition but neither with the predictions of the traditional RW model nor with the predictions of the modified RW model. Indeed, across the three rating phases, the change in participants’ ratings of objects A-D was consistent with the predictions of a simple Bayesian model applied to the IS and BB conditions.

**Discussion**

The goal of Experiment 3 was to determine how adults evaluate causally four objects and to determine to what extent their causal ratings were consistent with the predictions of a Bayesian model or one of two associative models. The results from Experiment 3 extended those from Experiment 2 by showing that adults’ causal ratings continued to be consistent with the predictions of the Bayesian model for both the BB and IS conditions and that non-acted on objects do not affect adults’ assessments of acted-on objects. The fact that adults' causal ratings in Experiments 2 and 3 conformed with the predictions of the Bayesian model but not with the predictions of the two associative models suggests that a Bayesian-inference mechanism underpins adults’ causal reasoning about four objects.

**General Discussion**

The main goals of the experiments were fourfold. First, the experiments were designed to examine whether adults engaged in BB and IS reasoning using pre-, mid-, and post-rating phases in three experiments. Second, the experiments investigated whether adults engaged in BB and IS reasoning in the context of two, three, and four objects. Recall that previous investigations that used the blicket-detector or blicket-detector-like paradigm examined only to what extent infants (Sobel & Kirkham, 2006), children (e.g., Gopnik et al., 2001; Sobel et al., 2004), and adults (e.g., Griffiths et al., 2011) engaged in BB reasoning in the context of two objects, and only the latter of these studies investigated BB and IS reasoning in adults across three different rating phases with two objects. Third, to our knowledge, this was the first series of studies to assess both IS and BB reasoning directly by comparing the extent to which adults’ causal ratings of two (Experiment 1), three (Experiment 2), and four (Experiment 3) objects were consistent with the predictions of one of three primary competing computational and analytical models. The fourth goal was to determine what role, if any, non-acted-on objects affect the causal ratings of acted-on objects given that Sobel and Kirkham (2006) observed such a relation with 24-month-olds.

In conjunction, the present data suggest that adults use associative processes to reason about two objects and Bayesian processes to reason about three and four objects. This is based on the fact that adults' causal responses were consistent with the predictions of two associative models in Experiment 1 and were consistent with the simple Bayesian model in Experiments 2 and 3: there was no discernible drop in adults' causal ratings of object B between any two of the three rating phases in Experiment 1, whereas adults' causal ratings of object B did undergo a drop between the mid- and post-rating phases in Experiments 2 and 3.

An outstanding question based on the present data is why the adults might have processed the events associatively in the context of two objects but in terms of a Bayesian-inference mechanism or process in the context of three and four objects. One plausible, albeit speculative, explanation is that when asked to reason about two objects, but not when asked to reason about three or four objects, adults may have abstained from engaging in effortful reasoning. This is because reasoning about two objects presumably recruits fewer cognitive resources than reasoning about three and four objects. This, in turn, may have caused or encouraged participants to process the events associatively when asked to reason about two objects. In line with this argument, it has been suggested that associative processing is thought generally to be an automatic, non-effortful process that yields a fast response, one that consumes few cognitive resources, and one in which the benefits of employing such a process outweighs the costs of employing a more effortful one (e.g., Coenen, Rehder, & Gureckis, 2015; Darlow & Sloman, 2010; Kahneman & Frederick, 2002; Osman, 2004; Smith & DeCoster, 2000). Given these arguments, it might well be predicted that when asked to reason about two objects—which presumably recruits few cognitive resources—adults would engage in associative processing.

Perhaps somewhat counterintuitively, however, because reasoning about three and four objects recruits more effortful and explicit processes, inadvertently adults may have been encouraged to use a more explicit and effortful process in Experiments 2 and 3 such as Bayesian inference. Note that the reason adults may have been encouraged to engage in more effortful and explicit processes presumably was because it becomes more difficult to rely on automatic and associative processes to track and disambiguate between causes and noncauses as the number of candidate causes increases. Thus, reasoning about two objects may have encouraged adults to expend fewer cognitive resources or effort and to provide responses that accorded with the predictions of the two associative models, whereas reasoning about three and four objects may have encouraged adults to engage in more effortful processing and thus provide responses that were consistent with the predictions of a simple Bayesian model. Note that such an account predicts that the size of the BB effect should have increased between Experiments 1 and 3. Although the magnitude of the BB effect did not differ reliably in Experiment 1 (*M* = 8.54), Experiment 2 (*M* = 12.5), and Experiment 3 (*M* = 13.15) perhaps because of statistical noise, it is clear from the pattern of the means that there was a tendency for the BB effect to increase across the three experiments. Nonetheless, future research would benefit by exploring this issue in more detail.

**How the present study extends extant research on BB reasoning**

This is the first series of experiments to use the blicket-detector task to show that adults engaged in Bayesian inference when asked to reason about three and four objects but not when asked to reason about two objects: the results revealed that adults processed the causal events associatively based on the traditional and modified RW models when asked to reason about two objects but in terms of a Bayesian-inference mechanism when asked to reason about three and four objects. The finding from Experiment 1—that adults may use something like the traditional RW model—is particularly striking given that it has been suggested that the traditional RW model cannot account for certain aspects of human causal reasoning such as BB and IS reasoning. This is because this model fails to account for the finding that learners treat object B differently between the BB and IS conditions (e.g., Sobel et al., 2004), that object B undergoes a drop between different rating phases in a blicket-detector-like task (e.g., Griffiths et al., 2011), that humans engage in blocking when cues are said to be causes than when they are said to be effects (Kloos & Sloutsky, 2013; Waldmann & Holyoak, 1992) or predictors (e.g., De Houwer, Beckers, & Glautier, 2002), and that humans engage in BB when causes are viewed submaximally such that two candidate causes do not combine to produce a larger effect than one of the two causes in isolation (Beckers, De Houwer, Pineno, & Miller, 2005; Lovibond et al., 2003). Yet the present data demonstrate clearly that such a model can account for adult causal reasoning given that adults did not engage in BB reasoning when they were asked to reason about two objects.

These experiments are important because they are the first to address key but until now unanswered questions in the causal literature. Specifically, these experiments were designed not only to determine to what extent adults engage in BB and IS reasoning but also to address whether BB reasoning depends jointly on the number of objects about which learners are asked to reason and on being asked to provide three sets of causal ratings. These experiments also extend previous findings because they rely on a direct technique to assess BB and IS reasoning: BB and IS reasoning were compared within, rather than between, conditions for both the BB and IS conditions, thereby eschewing cross-condition contamination. Recall that this approach differs fundamentally from previous research that based the claim that humans engaged in BB or IS reasoning on comparing the number or proportion of participants who chose the causally redundant cue across disparate conditions that should elicit different causal responses. Furthermore, these experiments, and especially the computational and analytical models presented throughout this paper, are meaningful in helping to elucidate the mechanism by which adults reason about causal events. With a small number of exceptions (e.g., Carroll, Cheng, & Lu, 2010; Griffiths et al., 2011), there have been few attempts to compare the causal ratings of adults in a BB and IS task to the formal predictions of analytical and computational models.

Finally, these experiments also represent a theoretical contribution to studies of BB reasoning. In particular we argue that when BB is taken to represent a drop in the rating of causally redundant cues between any two phases, BB is a phenomenon that can be accounted for by Bayesian and associative models alike. This represents an important extension to previous theorizing on this topic that maintained that associative models cannot account for BB reasoning presumably because models such as the traditional RW model fail to account for the differential treatment of object B across the BB and IS conditions (e.g., Sobel et al., 2004) or the observed drop in rating of object B between a mid- and post-rating phase (e.g., Griffiths et al., 2011), among other things. What we have demonstrated is that, at least in the context of two objects, adults do not treat objects A and B differently when a different metric is used to assess BB and IS reasoning. Indeed, when formulated as a drop in the rating of a redundant cause between any two rating phases, it is clear that both the modified RW model and a simple Bayesian model predict BB reasoning.

**Limitations of the present research**

Before closing, it is worth noting some potential criticisms of the present experiments. First, the experiments described here were designed to address an important limitation of previous BB and IS studies with children (Sobel et al., 2004) and infants (e.g., Sobel & Kirkham, 2007); that is, in contrast to previous studies with children that only indirectly examined BB and IS reasoning, these experiments were designed directly to examine both forms of reasoning. However, the experiments reported here do not directly address to what extent children engage in BB or IS reasoning because adults were tested rather than children. Given that the experiments reported here did not test children, it remains an open question whether children would show a similar pattern of responding that the adults showed in the present experiments. To the extent that one is interested in providing a developmental account of causal processing and reasoning, it will be important to test older infants and young children, in addition to testing adults. Critically, to allow direct comparison, these tasks should preserve the logic of the experiments reported here and, more importantly, the logic of the BB and IS event. The experiments reported here nonetheless represent an important extension to the extant body of research on this topic because they provide insight into how adults reason about the causal events used here that consisted of two, three, and four objects, about whether adults can engage in BB, and about the nature of the causal mechanism that underlies this ability and that can explain BB reasoning.

Second, although we compared the predictions of a simple Bayesian model, the traditional RW model, and the modified RW model applied to two, three, and four objects to adults' causal ratings of those objects, we did not consider to what extent adults' causal ratings were consistent with other plausible models of human causal reasoning such as the Power PC model (e.g., Cheng, 1997) or the traditional contingency model (Jenkins & Ward, 1965). Though it would have been ideal exhaustively to compare the predictions of these models to the causal ratings of the adult participants obtained here to determine the extent to which each model fit the human data, we chose instead to focus on the aforementioned three models because a large debate exists concerning whether, and to what extent, an associative or Bayesian mechanism underlies human causal learning. We also chose not to focus on models such as the Power PC model—although this model too has been discussed extensively in the context of human causal reasoning (for an extensive review see Cheng, 1997)—because this model does not make predictions for certain phases of the tasks used in Experiments 1-3. For example, in the context of the present experiment, the Power PC model makes no prediction about the pre-ratings for any of the objects in any of the conditions precisely because the relevant probabilities needed to compute each object's causal power are not yet known at this point. It is interesting to note that this is a limitation of contrast models more generally because the contrast between conditional likelihoods can only be computed once all relevant conditional probabilities are known. Nonetheless, an important goal for future research will be to compare the causal-reasoning performance of adults in a BB task to the predictions of a larger set of computational and analytical models than that used in the present study.

Finally, although we assessed adults' BB and IS performance using more objects than that typically used in previous research on this topic, a shortcoming of the present investigation is that it is unclear to what extent adults' causal-reasoning performance generalizes to novel contexts outside that used here. Indeed, although the specifics of particular studies on human causal learning differ, an overarching goal of studies of human causal learning is better to understand how and in what ways causal learning operates in the real world as well as to understand the mechanisms that support this process. Despite the fact that this is an issue that is not limited to the present study, future research would benefit by studying human causal learning in contexts that better approximate learning in the real world.

**Conclusion**

These potential criticisms notwithstanding, these experiments constitute one of the first systematic attempts to examine BB and IS reasoning in the context of multiple objects, different contexts, and across a series of rating phases. It has been previously proposed that humans use Bayes' rule to reason about causal events and that the putative BB and IS findings can only be accounted for by a Bayesian inference mechanism (e.g., Gopnik et al., 2004; Sobel et al., 2004). The experiments reported here support a different conclusion. These experiments demonstrate that adults use both associative processes and Bayesian inference to process causal events, and the process that is employed depends on the number of objects about which adults are asked to reason.

**ACKNOWLEDGEMENT**

We thank Kaily Bruch and Marina Selenica for their help with data collection and participant recruitment. We also thank the participants who kindly agreed to participate in the research.This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *34*(1), 1.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*(4), 381-405.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 238.

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General*, *135*(1), 92.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in cognitive sciences*, *18*(10), 497-500.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. *The developmental psychology of time*, 209-254.

Carroll, C. D., Cheng, P. W., & Lu, H. (2013). Inferential dependencies in causal inference: A comparison of belief-distribution and associative approaches. *Journal of Experimental Psychology: General*, *142*(3), 845.

Chapman, G.B. (1991). Trial-order affects cue inte raction in contingency judgement. J ourna l of Experi- menta l Psychology : Lea rning, Memory, a nd Cognition, 17, 837-854.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, *18*(5), 537-545.

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological review*, *104*(2), 367.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology*, *79*, 102-133.

Darlow, A. L., & Sloman, S. A. (2010). Two systems of reasoning: Architecture and relation to emotion. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(3), 382-392.

Dickinson, A. (1996). Within compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology: Section B*, *49*(1), 60-80.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, *37*(5), 620.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, *111*(1), 3.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407-1455.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233-259.

Houwer, J. D., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(3), 965-985.

Jeffreys, H. (1961). Theory of probability (3rd ed.). Oxford: Oxford University Press, Clarendon Press

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological monographs: General and applied*, *79*(1), 1.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, *49*, 81.

Kloos, H., & Sloutsky, V. M. (2013). Blocking a redundant cue: what does it say about preschoolers' causal competence?. *Developmental science*, *16*(5), 713-727.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636-645.

Larkin, M. J., Aitken, M. R., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1331.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality?. *Cognition*, *25*(3), 265-288.

Lovibond, P. F, Been, S., Mitchell, S.J., Bouton, M.E., Fronhardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory and Cognition*, *29*(1), 97.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284-299.

Mayrhofer, R., & Waldmann, M. R. (2016). Sufficiency and necessity assumptions in causal structure induction. *Cognitive science*, *40*(8), 2137-2150.

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: observational causal learning in 24-month-old infants. *Developmental psychology*, *48*(5), 1215.

Morey, R.D, & Rouder, J.N. (2018), BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.2.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development*, *5*(2), 193-207.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review*, *11*(6), 988-1010.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186-216.

R Development Core Team (2008). R: a language and environ- ment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111-163.

Rakison, D. H. (2005). A secret agent? How infants learn about the identity of objects in a causal scene. *Journal of Experimental Child Psychology*, *91*(4), 271-296.

Rakison, D. H., Smith, G. T., & Ali, A. (2016). Who is the dynamic duo? How infants learn about the identity of objects in a causal chain. *Developmental psychology*, *52*(3), 355.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, *72*, 54-107.

Rehder, B. (2018). Beyond Markov: Accounting for independence violations in causal reasoning. *Cognitive psychology*, *103*, 42-84.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64-99.

Rothe, A., Deverett, B., Mayrhofer, R., & Kemp, C. (2018). Successful structure learning from observational data. *Cognition*, *179*, 266-297.

Schlottmann, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *The Quarterly Journal of Experimental Psychology*, *44*(2), 321-342.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1-21.

Shanks, D. R. & Dickinson, A. (1987). Associative accounts of causality judgment. In Bower, G. H. (Ed.), The psychology of learning and motivation, Vol. 21. San Diego: Academic Press.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and social psychology review*, *4*(2), 108-131.

Sobel, D. M. (2004). Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals. *British Journal of Developmental Psychology*, *22*(1), 37-58.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental psychology*, *42*(6), 1103.

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants’ developing statistical reasoning abilities and their representation of causal knowledge. *Developmental science*, *10*(3), 298-306.

Sobel, D. M., & Munro, S. (2006). When Mr Blicket wants it, children are Bayesian. In *Proceedings of the Cognitive Science Society* (pp. 810-816).

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive science*, *28*(3), 303-333.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, *25*(2), 127-151.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological science*, *25*(1), 161-169.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child development*, *88*(1), 229-246.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 694.

White, P. A. (2006). How well is causal structure inferred from cooccurrence information?. *European Journal of Cognitive Psychology*, *18*(03), 454-480.

Figure Captions

Figure 1. The four different causal graphical models indicating the possible causal relations for a causal event that involves two objects and one blicket detector. *A* and *B* correspond to the two objects that were used on the machine and *E* indicates the activation of the machine.

Figure 2. Participants’ causal ratings of objects *A* and *B* before each of the four test event sequences were shown, midway through each event sequence, and after the entire event sequence had been demonstrated.

Figure 3. Participants’ causal ratings of objects *A*, *B*, and *C* before each of the four test event sequences were shown, midway through each event sequence, and after the entire event sequence had been demonstrated.

Figure 4. The eight different causal graphical models indicating the possible causal relations for a causal event that involves three objects and one blicket detector. *A*, *B*, and *C* correspond to the three objects that were used on the machine and *E* indicates the activation of the machine.

Figure 5. Participants’ causal ratings of objects *A*, *B*, *C*, and *D* before each of the four test event sequences were shown, midway through each event sequence, and after the entire event sequence had been demonstrated.

Figure 6. The sixteen different causal graphical models indicating the possible causal relations for a causal event that involves three objects and one blicket detector. *A*, *B*, *C*, and *D* correspond to the fourt objects that were used on the machine and *E* indicates the activation of the machine.

Appendix

**The Bayesian model**

A fundamental assumption of the Bayesian inference approach is that causal induction is a process that involves representing the entire space of candidate causal hypotheses—which are expressed as parameterized causal graphical models with nodes that are connected by edges that encode the Markov condition—and then choosing the hypothesis that is most consistent with the data by application of Bayes' rule. Formally, it is assumed that, at the beginning of a task, experiment, or learning episode, an ideal Bayesian learner represents all possible candidate hypotheses, *H*, whereby each hypothesis, *h* ∈ *H*, is assigned some prior probability, *p*(*h*). This prior probability represents the learners’ confidence that the observed data were generated by a given causal hypothesis. Following observations of data, *d*, the learner then uses Bayes' rule to compute and assign posterior probabilities to each hypothesis, *p*(*h*|*d*),

,

where *p*(*d*|*h*) represents the likelihood or the probability of the data *d* under a given hypothesis *h* normalized by members of the set. Because the hypotheses in all of the experiments presented here are assumed to be deterministic (i.e., objects either produce or do not produce detector activation), the likelihoods are set to 1 whenever a link (i.e., causal relation) exists in the hypothesis (Figure 1) and is consistent with the observed data; otherwise, they are set to 0.

The first step in defining a model of the task presented in Experiments 1 to 3 was to specify the hypothesis space *H* and the hypotheses *h* that comprise that space. This step is necessary before Bayes' rule can be used to determine the hypothesis with the largest posterior probability. Because Experiment 1 uses two objects (i.e., two candidate causes), Experiment 2 uses three objects (i.e., three candidate causes), and Experiment 3 uses four objects (i.e., four candidate causes), the hypothesis space for each experiment, respectively, consists of four, eight, and 16 hypotheses (see Figures 3, 5, and 7 above). The specific parameterization of each hypothesis in the space is specified by the activation law, which, for all three experiments, states that the blicket detector will activate if, and only if, a blicket object contacts it. The second step in defining this model is to specify the prior probabilities of each hypothesis in each of the three experiments. If we assume that the probability that a particular object is a blicket is independent of the probability that other objects are blickets, then prior probabilities for each of the three experiments for both the BB and IS conditions can be found in Tables 2, 5, 8, 11, 14, and 17. These prior probabilities are then used to compute posterior probabilities for each hypothesis when new data is observed according to Bayes' rule.

Because we are using deterministic hypotheses—which is not a necessary requirement (for an extended discussion about deterministic and non-deterministic hypotheses, see Griffiths et al., 2011)—whenever a link exists the likelihood is set to 1 and 0 otherwise. Once we have determined whether such a link exists for a particular object, we can compute the likelihood that object A or B, A, B, or C, or A, B, C, or D is a blicket in Experiments 1 to 3, respectively, by taking the product of the likelihood that a particular object activated the detector under each hypothesis (according to the aforementioned procedure) and the prior probability of each hypothesis and then summing this product. To determine the probability that object B is a blicket, for example, we can compute the following equation:

,

where is 1 if a causal link between *B* and *E* existsfor a specific hypothesis *h*; otherwise, is 0. As is mentioned above, the qualitative predictions of this model for how participants should rate each of the objects in Experiments 1 to 3 are shown in Tables 2, 5, 8, 11, 14, and 17.

**Associative models**

*Rescorla-Wagner model*. In contrast to the Bayesian model, the traditional RW model (Rescorla & Wagner, 1971) uses error correction to produce causal ratings of the test events. On this model, error correction is computed according to the following equations:

∆*Vi* = α*i*β1(λ – ΣVk)

and

*V*n+1= *Vi* + ∆*Vi* ,

where ∆*Vi* is the change in the associative weight between the candidate cause *i* and effect *e*, α is the salience of the candidate cause *i*, β is the salience of the effect *e*, V is the current associative strength between the candidate cause *i* and effect *e*, and ΣVk is the summed associative strength between all candidate causes *i* to *k* and the effect *e*. Unlike the Bayesian model presented above —which did not require explicit simulation and thus represented an analytical model—for the RW model we simulated all three experiments to examine what predictions were generated and to determine whether those predictions differed from those of the Bayesian model for these experiments across the three rating phases. Note that because the RW model does not inherently espouse a parameter that accounts for participants’ baseline beliefs about causes (prior to the experiment) we set the weight for A and B in all three experiments to equal .5 to reflect that fact participants’ pre-ratings of both objects across all three experiments did not differ reliably from 50%; that is, participants seemed implicitly to assume that each object possessed an equal chance of being a cause prior to the presentation of any of the condition events in all three experiments. Finally, the model received 20 AB+ training trials and 20 A+ training trials. Because adults in Experiments 1-3 were asked to provide three sets of ratings—namely, pre-, mid-, and post-ratings—the model’s predictions at training trial 20 were used as the midway ratings of this model. Likewise, training trial 40 was taken to be the model’s post-ratings for all four objects across the three experiments. The model’s pre-ratings were assessed before initiation of the simulation. Following these trials, A and B were presented alone and the model’s prediction about the causal status of both objects was assessed in Experiment 1 (or the model’s predictions about each of three or four objects were assessed in Experiments 2 and 3, respectively). Note that object C did not participate in the simulation of either the AB+ trials or the A+ trials given that this object in Experiment 2 and objects C and D in Experiment 3 were never shown alone on the detector. Note also that because we were concerned with the qualitative predictions of the model rather than with the extent to which participants’ causal ratings matched the quantitative predictions of the model and because only the qualitative predictions are presented for the Bayesian model, we only report below the qualitative predictions of the RW model for Experiments 1-3. This approach is also consistent with that taken in Griffith et al. (2011). The traditional RW model predictions for all three experiments can be found in Tables 3, 6, 9, 12, 15, and 18.

*Modified Rescorla-Wagner model.* Despite the fact that the form of the learning rule for the RW model and the modified RW model is identical (i.e., ∆*Vi* = α*i*β1(λ – ΣVk), the two models differ in their treatment of the salience parameter, α. Whereas α is set to 0 for absent cues in the traditional RW model because such cues are assumed to have no salience, α is set to a nonzero (negative) value because absent cues are assumed to be negatively correlated with the outcome in the modified RW model. Besides the difference in how both models treat α, the training procedure used for the modified RW model was identical to that used for the traditional RW model. The predictions of the modified RW model for all three experiments can be found in Tables 4, 7, 10, 13, 16, and 19.

Figure 1



Figure 2

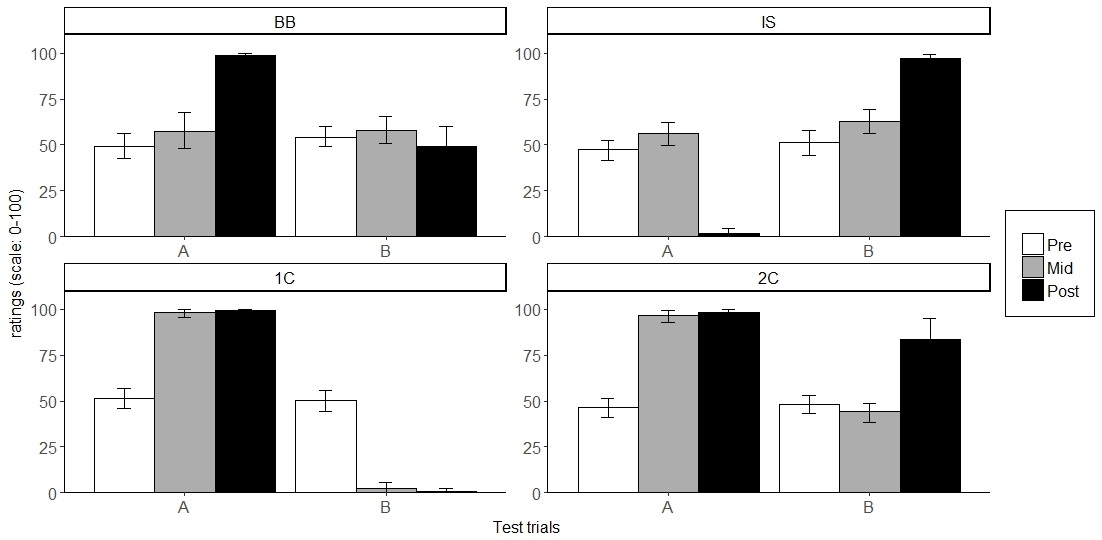


Figure 3

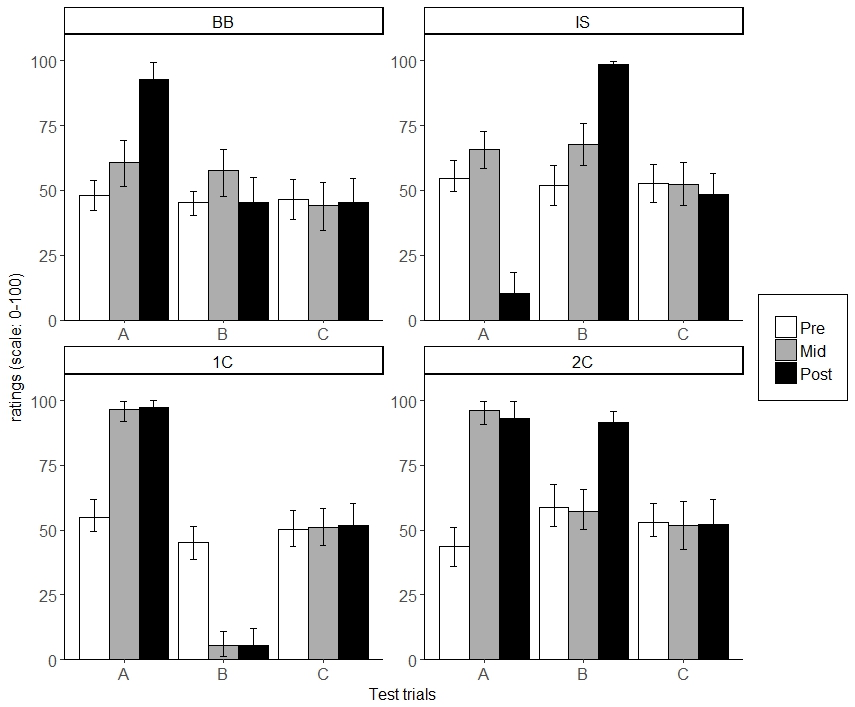


Figure 4

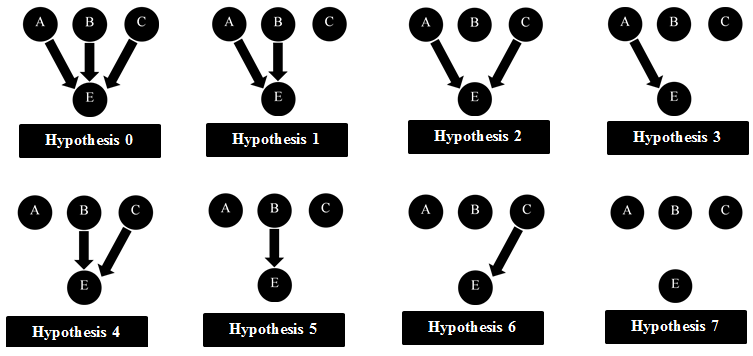
****

Figure 5

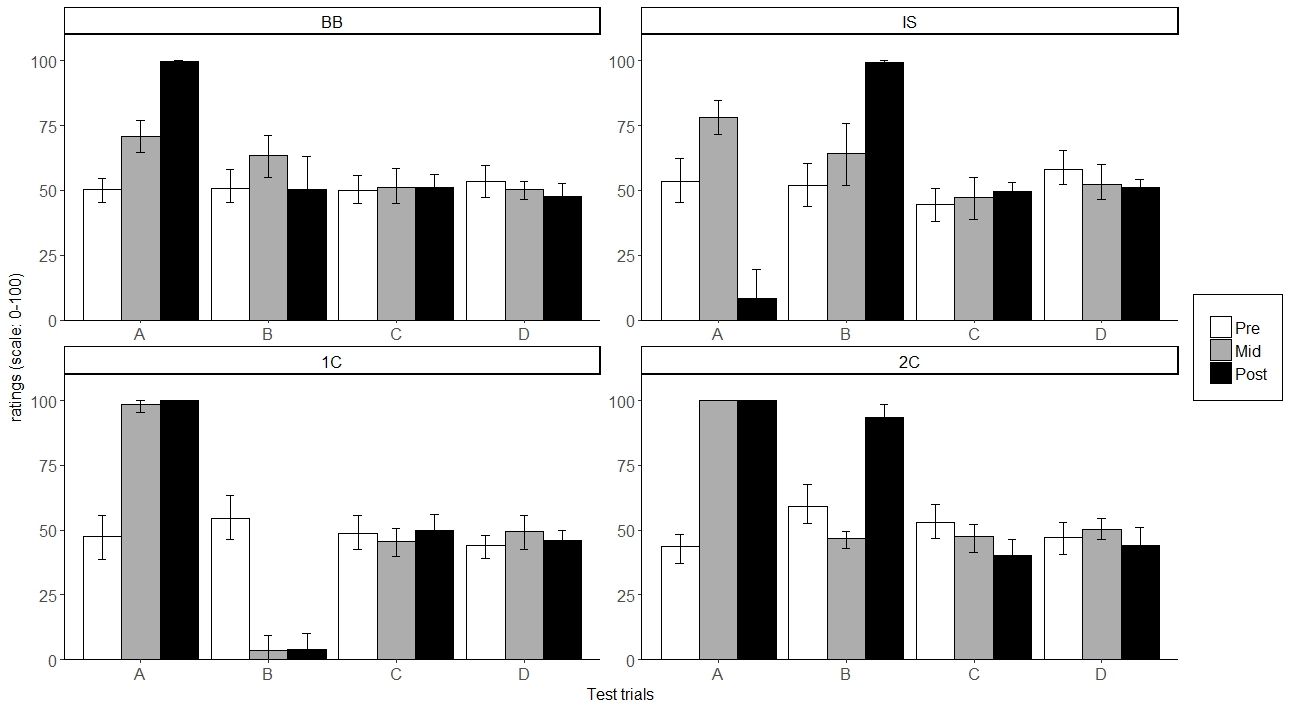


Figure 6

