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 human reasoners to learn about the complex interactions in the world around them. However, the available evidence with children and adults suggests that the mechanism or set of mechanisms that underpins causal perception and 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 the backward-blocking finding in children and adults (e.g., Sobel, Tenenbaum, & Gopnik, 2004). However, the evidence is mixed about whether, and to what extent, adult reasoners engage in BB reasoning. Here, we report four experiments that examine to what extent adults engage in BB reasoning and what mechanism best explains their behavior in the task.

There is perhaps no ability that is more important 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 general 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., Gopnik et al., 2001; Sobel & Kirkham, 2006; Sobel et al., 2004; 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. This research has demonstrated that starting at 18 months of age children can use higher-order relational rules to make causal inferences (Benton, Rakison, & Sobel, 2021; Walker & Gopnik, 2014). By 2 years of age children not only can categorize and distinguish blickets from non-blickets but they can use patterns of variation and covariation to make causal inferences and to generate causal interventions (e.g., Sobel & Gopnik, 2000; Gopnik et al., 2001). In addition, 2-year-olds can design causal interventions based on the observed interventions and actions of others (e.g., Meltzoff, Waismeyer, & Gopnik, 2012). By 3 years of age children show sensitivity to base-rate information in their causal inferences about blickets (Sobel, Tenenbaum, & Gopnik, 2004; Sobel & Munro, 2006), and between 3 and 5 years of age they can reason about causes that span multiple domains such as the domains of psychology and biology.

One finding that has garnered significant attention in the developmental literature—and one that bears significantly on the question that motivates the present study, which concerns what the cognitive mechanism is by which children reason about causal events—is that by Sobel et al.'s (2004; see Griffiths et al. 2011 for related findings). These authors found that 4-year-old children—and to a lesser extent 3-year-old children—can engage in two forms of retrospective reevaluation, which refers to the process whereby learners use current causal information to reevaluate past causal information. The two forms of retrospective reevaluation that were investigated were backwards-blocking reasoning (henceforth, BB reasoning) and indirect screening-off reasoning (henceforth, ISO 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 retrospectively to disregard or block 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 from being considered to be a cause when A alone fails to activate the detector.

Sobel et al. (2004) found that 4-year-olds—and to a lesser extent, 3-year-olds—indicated through their verbal responses and intervention actions that object B, but not object A, was the blicket in the ISO condition, whereas the same-age children in the BB condition indicated that object A, but not object B, was a blicket. Both patterns of responses would be expected if children engaged in retrospective reevaluation. Crucially, 4-year-olds treated object A from the ISO trials as less of a blicket than object B from the BB trials. This finding was interpreted as evidence for BB reasoning. 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 reasoning—where a difference in treatment of object A between the BB and ISO conditions was used as evidence for such reasoning--and could also use base-rate information if the activation of the detector was presented as machine with desires rather than a generic machine that worked mechanically. In other words, if eyes were appended to the machine and it was called “Mr. Blicket”, 3-year-olds showed evidence of BB reasoning. These findings were interpreted to mean that a Bayesian-inference mechanism rather than an associative-learning mechanism underlies children’s BB reasoning performance. This is based on the fact that these findings were better predicted and explained by a Bayesian-inference mechanism but not by an associative-learning mechanism, which in this case was the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972). The RW model erroneously predicts that object A should be treated equivalently across the BB and ISO condition, whereas a Bayesian-inference mechanism predicts that A should be treated as less of a cause in the ISO condition compared to the BB condition. Note that the performance of the RW model was considered erroneous because it made a prediction that did not comport with participants’ actual responses (see below for why the model predicts that participants should treat A equivalently across the conditions). However, treating as equivalent object A across both conditions is a prediction that derives naturally from the mathematics that underlie the RW model.Despite the fact that Griffiths et al. (2011) interpreted these findings to mean that adults engaged in BB reasoning, the conclusion that Bayesian inference underlies children’s BB reasoning performance should be interpreted with caution. This is because BB has been given at least four distinct interpretations in the research literature. For example, some have interpreted BB to mean that object B undergoes a categorical drop between the initial and final rating phases presumably because A is shown explicitly to produce the effect during the A+ learning trial (Van Hamme & Wasserman, 1983). In contrast, others have argued—based on a Bayesian analysis—that BB refers to a return to baseline in the rating of B between the two rating phases (Griffiths et al., 2011). Still others have interpreted the differential treatment of object B between the BB and IS conditions as evidence for BB reasoning; that is, learners are said to engage in BB reasoning if they are more likely to choose (and intervene on) object B in the IS condition than in the BB condition (e.g., Sobel et al., 2004). A final group of researchers have argued that BB refers to the difference in treatment of object B in the BB condition to that in a control condition in which neither object A nor object B is shown to produce the effect by itself; that is, learners are said to engage in BB reasoning if they are more likely to choose object B in a control condition (i.e., AB+ only) than in the BB condition (AB+ A+; Larkin, Aitken, & Dickinson, 1998; Lovibond et al., 2003; Shanks, 1985). Given these competing interpretations of BB reasoning, one must exercise caution in asserting that participants engaged in BB reasoning in Griffith et al.’s (2011) study. return to this issue in more detail in the Discussion following Experiment 1.

**Bayesian inference as an account of BB reasoning**

Nonetheless, the fact that children evaluated object B differently between the BB and IS conditions in Sobel et al. (2004) and that adults engage in BB reasoning and are sensitive to base-rate information in Griffiths et al. (2011) have been interpreted proponents of the Bayesian-inference perspective to mean that humans use a simple form of Bayes’ rule rather than associative learning to reason about causal events. The crux of this account is that human learners use a simple form of Bayes’ rule to reason about causal events and to choose the hypothesis that is most consistent with the observed data within a space that consists of multiple competing causal hypotheses (e.g., Sobel et al., 2004; Sobel & Wellman, 2012). These findings have been argued to challenge certain rudimentary associative models such as the Rescorla-Wagner model (henceforth, RW; Resorla & Wagner, 1972) and certain contingency models such as the Power PC model (e.g., Cheng, 1997). These findings challenge the RW model because it predicts that B should be treated equivalently across the BB and IS conditions. This is because this model only makes weighted adjustments to cues that are present and not to cues that are absent, which B is not during the "A" phases in both the BB and IS conditions. In other words, given that B is absent during the A phases of the BB and IS tasks, the RW model (Rescorla & Wagner, 1972) predicts that the associative strength between B and the blicket effect should remain unchanged across the experimental trials in both conditions.

The reason these findings challenge both associative- and contingency-based models is because they require many learning trials for reliable associations to be established, in contrast to the relatively small number of learning trials that the children and adults required in Sobel et al.'s (2004) and Griffiths et al.’s (2011) studies to make causal inferences. Finally, the BB and IS findings challenge both associative- and contingency-based models—including more sophisticated associative- and contingency-based models such as those of Van Hamme and Wasserman (1994),Dickinson and Burke (1996), and the contingency model of Cheng (1997), all of which can account for BB and IS—because these models do not naturally encode base rates and thereby have no way to account for the findings that children (e.g., Bonawitz & Lombrozo, 2012; Griffiths et al., 2011; Sobel & Munro, 2006; Sobel et al., 2004) and adults (e.g., Griffiths et al. 2011) incorporate base rates into their causal decisions.

Despite these findings, open questions remain about whether, and to what extent, humans engage in backwards BB. This is because the conclusion that children in Sobel et al. (2004) engaged in backwards BB, for example, was based on a comparison of the proportion of children who chose B in the BB condition to the proportion of children who chose B in the IS condition; that is, 4-year-olds were said to have engaged in BB reasoning because fewer of them chose B as a cause in the BB condition than in the IS condition. However, we argue that this comparison may not be the most appropriate one to assess BB reasoning. This is because, as discussed above, BB reasoning implies either a categorical drop in the rating of B within the BB condition itself or a return to baseline within the same condition. Thus, the more appropriate comparison—which was not undertaken in Sobel et al. (2004) but was in Griffiths et al. (2011)—would have been to compare the proportion of children who chose B before the BB event was presented or adults’ causal ratings of both objects before the BB event to both sets of responses after the event was demonstrated.

In addition to interpretational issues, closer inspection of the literature reveals that the evidence is mixed about whether, under what conditions, and to what extent humans engage in BB reasoning. In particular, some studies purport to show evidence of BB reasoning in children and adults (e.g., Griffiths et al., 2011; Sobel et al., 2004), whereas other studies either show no evidence or only weak evidence of BB (e.g., Larkin, Aitken, & Dickinson, 1998) or showed evidence of BB reasoning under some conditions but not under others (e.g., Lovibond et al., 2003). For example, Lovibond et al. (2003) demonstrated that if participants learned first that foods combined additively to produce larger allergic reactions—that is, two foods could potentially combine to produce a larger allergic reaction than either of the food cues could produce in isolation—and then were shown the BB trial, they tended to rate food B as less of a cause than the ratings of foods C and D in a control condition. Note that neither food C nor food D in the control condition was shown in isolation; rather, both food cues together produced the reaction. In contrast, if participants were taught that it is not possible to determine outcome additivity perhaps because of an imposed ceiling on the magnitude of allergic reactions, participants' ratings of food B did not differ from those of C or D in the control condition. This finding is important because it demonstrates that BB reasoning depends on the extent to which participants are made to think about causes additively or in terms of an impost ceiling causes are viewed additively and perhaps is not as robust as originally thought (for related findings, see also Beckers, de Houwer, Pineno, & Miller, 2005).

Finally, although BB reasoning typically has been assessed in the context of two objects, to date only one study by Sobel and Kirkham (2006) has examined BB and IS reasoning in the context of three or more objects. In this study, an experimenter first placed three objects (i.e., objects A, B, and C) on a table and then showed 19- and 24-month-old children that objects A and C together made the detector activate when they were placed on it. Object C was then placed alone on the detector, which either caused the detector to activate (BB condition) or not (IS condition). Sobel and Kirkham (2006) reasoned that if participants engaged in BB reasoning and subsequently blocked object A, then they should be more likely to use object B to make the detector activate than A despite the fact that object B was never demonstrated alone on the machine. The results revealed that when the 19- and 24-month-olds were subsequently given objects A and B and were asked to make the machine go, only the 24-month-olds were more likely to place object A on the detector in the IS condition than in the BB condition. The 19-month-olds’ performed at chance. Sobel and Kirkham (2006) interpreted the 24-month-olds’ categorization behavior to mean that they engaged in BBBB reasoning. However, as was discussed above, this may be not the appropriate comparison to assess BB reasoning; that is, because the drop in the proportion of A choices within the BB condition was not assessed, it cannot be concluded definitively that children engaged in BB reasoning. It remains an open question, then, whether, under what conditions, and to what extent learners engage in BB reasoning for two or more objects.

**Present experiments**

In light of the issues outlined above, the aims of Experiments 1-5 were threefold. The first aim of the experiments—based on the mixed evidence for BB reasoning—was to replicate Sobel et al.’s (2004) study to determine whether adults engage in BB reasoning. The reason we tested adults is because neither adults' causal-reasoning abilities nor their ability to engage in BB reasoning have been tested in the original blicket-detector task; rather, adults’ ability to engage in BB reasoning has either been assessed in variants of the blicket detector task (e.g., Griffiths et al., 2011) or in tasks that are entirely unrelated to the blicket-detector task (e.g., Shanks, 1985; Shanks & Dickinson, 1987; Lovibond et al., 2003).

A second and related aim of the experiments was to implement pre- and post-rating phases in Experiments 1 and pre-, mid-, and post-rating measures in Experiments 2-5 to determine whether and to what extent participants’ ratings of the redundant cause B undergoes a drop between each rating phase. This method of assessing BB represents an important extension to previous studies because implementing multiple rating phases makes it possible to track how B changes across the multiple ratings phases within the BB condition itself—without recourse to comparisons of B in two disparate conditions that should produce disparate evaluations of B–and because it possible to test the predictions of multiple competing models to determine the mechanism that may underlie adults’ causal-reasoning abilities.

The third and final aim of the experiments was to assess BB reasoning in the context of multiple objects. As has been discussed, with the exception of Sobel and Kirkham’s (2006) study, BB reasoning has not been assessed in the context of three or more objects. Thus, Experiments 1, 2 and 5 used two objects to assess BB reasoning in adults, whereas BB reasoning was assessed using three and four objects in Experiments 3 and 4, respectively.

In general, Experiments 1-4 use the blicket-detector task to assess causal reasoning in an adult sample. In Experiment 1, adults were asked to rate the likelihood that each of two objects (A and B) were blickets across two rating phases and in four conditions that included the BB and IS events. In Experiments 2-4, adults were asked to rate the likelihood that each of two (Experiment 2), three (Experiment 3), or four (Experiment 4) objects are blickets in four conditions that included the BB and IS events. Critically, in Experiments 2-4 adults were asked to rate the likelihood that the objects are blickets both before a particular trial (i.e., pre-ratings), midway through the trial (i.e., mid-ratings), as well as after the trial (i.e., post-ratings).

Experiment 5 used a novel procedure to assess causal reasoning in adults. Adults were shown video sequences that consisted of two circles and a centrally located box. The box contained a sun that “popped up” when one of the two objects contacted it. Participants were asked to determine which of the two objects caused the sun to pop up in the same four conditions as those used in Experiments 1-4, and again these ratings were assessed at three time points: before a particular event (pre-rating), after half of particular event had been shown (mid-rating), and after the full event had been shown (post-rating).

**Experiment 1**

**Methods**

**Participants.** Sixty college students were recruited from Carnegie Mellon University to participate in Experiment 1. There were approximately equal numbers of males and females in the final sample. Note that an a priori power analysis revealed that approximately 16 participants would be needed to have an 80% chance to detect a reliable difference if such a difference exists at α = .05 (Cohen, 1988). Thus, we can be confident that these experiments are sufficiently powered. Participants received credit for an introductory-level psychology course.

**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 remote control was pressed and the object that was predetermined to be a blicket was placed on the surface of the detector, the music and the lights began to play and 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.

The stimuli 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 using 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 that the machine 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 two pretest trials to ensure that they understood the purpose of the experiment and how the machine worked. In one of the pretest trials, one of two objects—which were unrelated to those used during the main experimental phase—activated the machine and was labeled a blicket, whereas the other of the two objects (randomly determined) did not activate the machine. All participants (N = 60) extended the blicket label to the correct causal object and thus were included in the main analysis, χ2(1) = 120, p < .00001.

Following the pretest phase, 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 BB (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 and were on task. 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. Both blocks were then placed together on the machine twice, which caused the detector to activate both times. 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. 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. 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. In each of the four conditions, participants were instructed to rate on a scale that ranged from 0 (definitely not) to 100 (definitely is) how confident they were that each object in the pair was a blicket both before and after a trial for a total of two causal ratings.

**Results**

All analyses were conducted in R (R Development Core Team, 2008). Figure 1 shows the mean ratings of the causal likelihood that objects A and B were blickets across the four conditions. 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. 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).[[1]](#footnote-1)

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

Insert Figure 1 about here

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

**Control conditions.** The first analysis examined the distribution of responses to the A and B objects in the 1C condition. A mixed-effects model was fit to participants' pre- and post-ratings of object A in the 1C condition. Participants were included as a random-effect factor and Causal Likelihood Rating (pre-rating of A vs. post-rating of A in the 1C condition) was included as the fixed-effects factor using simple dummy coding (0 = Pre-Rating of A, 1 = Post-rating of A)[[2]](#footnote-2). The model showed a significant main effect of Causal Rating, Type III *F*(1,59) = 234.56, *p*<.0001, which indicated that participants' post-ratings of A (Δβ = 45.22, *M* = 94.82, Bootstrapped 95% CI[39.39, 51.04])[[3]](#footnote-3) were significantly higher than their pre-ratings of A (β = 49.6, *M* = 49.6, Bootstrapped 95% CI[45.63, 53.57]) in the 1C condition. Additionally, a mixed-effects model was fit to participants' pre- and post-ratings of object B in the 1C condition to determine whether, as would be expected, participants' causal ratings of B dropped between the pre- and post-rating phases (because B would have been shown not to be a cause).

This model also showed a main-effect for Causal Rating, Type III *F*(1,59)=131.19, *p* < .0001, which indicated that participants provided lower post-ratings of B (Δβ = -39.07, *M* = 10.5, Bootstrapped 95% CI[-46.19,-31.94]) than the pre-ratings of it (β = 49.57, *M* = 49.57, Bootstrapped 95% CI[45.26, 53.88]. These analyses were corroborated by repeated-measures ANOVA with Causal Rating as the sole factor. This analysis revealed a significant main effect of Causal Rating, *F*(1, 59)=46.73, *p* < .0001, η2p = .44. Follow-up planned comparisons (with Bonferroni correction) using non-parametric Wilcoxon signed-ranked tests revealed that adults provided higher post-ratings of object A (*M* = 94.82) than pre-ratings of A (*M* = 49.6), *z* = 33, *p* < .0001, whereas they provided higher pre-ratings of object B (*M* = 49.57) than post-ratings of object B (*M* = 10.5), *z* = 1453.45, *p* < .0001.

The second analysis examined the distribution of responses to the A and B objects in the 2C condition. A mixed-effects model was fit to participants' pre- and post-ratings of object A and B in the 2C condition. The first mixed-effects model was fit to participants' pre- and post-ratings of object A in the 2C condition. The model showed a significant main effect for Causal Rating, Type III *F*(1,59) = 277.23, *p* <.0001, with participants providing higher causal post-ratings of A (Δβ = 44.82, *M* = 94.75, Bootstrapped 95% CI[39.38, 50.25]) than their causal pre-ratings of A (β = 49.93, *M* = 49.93, Bootstrapped 95% CI[45.95, 53.91]).

Likewise, a mixed-effects model was fit to participants' pre- and post-ratings of B in the 2C condition, which revealed a main effect of Causal Rating, Type III *F*(1,59) = 109.86, *p* < .0001. In particular, participants provided a higher post-rating of object B (Δβ = 29.6, *M* = 78.35, Bootstrapped 95% CI[23.52,35.68]) than their pre-rating of B (β = 48.75, *M* = 48.75, Bootstrapped 95% CI[23.52, 35.68]). These analyses were corroborated by repeated-measures ANOVA with Causal Rating as the sole factor. Because Mauchly's test indicated that the assumption of sphericity was violated, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .77), χ2(3) = .614, *p* < .0001. There was a main-effect of Causal Rating, *F*(3.89, 252.86) = 111.65, *p* < .001. Follow-up planned comparisons (with Bonferroni correction) using non-parametric Wilcoxon signed-ranked tests revealed that adults provided higher post-ratings of A (*M* = 94.75) than pre-ratings of object it (*M* = 49.93), *z* = 4, *p* < .001. Likewise, participants provided higher post-ratings of B (*M* = 78.35) than pre-ratings of it (*M* = 48.75), *z* = 74.5, *p* < .001. Together, the results from the 1C and 2C conditions presented here replicated those from Gopnik et al. (2001).

**IS condition.** We next examined whether adults engaged in IS reasoning. It was predicted that if participants engaged in IS reasoning, their pre-ratings of object A should be higher than their post-ratings of object A, whereas their pre-ratings of object B should be lower than their post-ratings of object B. Thus, to examine whether adults provided higher pre-ratings of A than post-ratings of A, we fit a mixed-effects model to their pre- and post-ratings. The model revealed a significant effect of Causal Rating, Type III *F*(1,59) = 255.59, *p* < .001, which confirmed that participants' post-ratings of A (Δβ = -43.17, *M* = 12.03, 95% Bootstrapped CI[-51.82, -34.51]) was significantly lower the pre-ratings of A (β = 55.2, *M* = 55.2, Bootstrapped 95% CI[50.15, 60.25]). We next fit a mixed-effects model to participants' pre- and post-ratings of B to determine if, after observing that A failed to produce the effect, participants provided higher post-ratings of B than pre-ratings of it. This model revealed a significant main-effect for Causal Rating, Type III *F*(1,59) = 184, *p* < .0001, which confirmed that participants' post-ratings of B (Δβ = 40.82, *M* = 90.32, 95% Bootstrapped CI[34.72,46.92]) were significantly higher than their pre-ratings of B (β = 49.5, *M* = 49.5, 95% Bootstrapped CI[45.41, 53.59]). Based on these results, it was concluded that adults engaged in IS reasoning.

**BB condition.** The next analysis examined whether adults engaged in BB reasoning. Participants who engaged in this form of reasoning were predicted to provide higher post-ratings of A than pre-ratings of it because A was shown to produce the effect in isolation, whereas their post-ratings of B were expected to be lower than their pre-ratings of B. Thus, we first fit a mixed-effects model that included participants’ pre- and post-ratings of A in the BB condition to determine whether, in fact, their post-ratings of A were higher than their pre-ratings of A. The model revealed a significant main effect of Causal Rating, Type III *F*(1,59) = 390.39, with participants providing higher post-ratings of A (Δβ = 45, *M* = 96.25, Bootstrapped 95% CI[40.4,49.59]) than pre-ratings of A (β = 51.25, *M* = 51.25, Bootstrapped 95% CI[48.03,54.47]).

The analysis of primary interest examined whether there was a drop in the rating of B between the pre- and post-phases. Recall that a drop in the rating of object B between the pre- and post-rating phases should be observed if participants engaged in BB reasoning. In particular, this analysis examined whether participants provided lower post-ratings of B than pre-ratings of it. Participants' pre- and post-ratings of B were thus entered into a mixed-effects model with Causal Ratings as the sole factor. Unlike the previous models that demonstrated robust effects, this model revealed a marginally significant main effect for Causal Rating, Type III *F*(1,59) = 237.44, *p* = .07, which demonstrates that participants' post-ratings of B (Δβ = -4.25, *M* = 42.25, Bootstrapped 95% CI[-9.34, 0.84]) were marginally lower than their pre-ratings of B (β = 46.5, *M* = 46.5, Bootstrapped 95% CI[43.46, 49.54]).

To determine whether this marginally significant difference was meaningful and provided evidence in support of the alternative hypothesis (i.e., that the mean pre- and post-ratings of B differed) rather than the null hypothesis (i.e., that the mean pre- and post-ratings of B did not differ), the data were examined next by estimating a Bayes factor using Bayesian Information Criteria (Wagenmakers, 2007). In this analysis, we compared the fit of the data under the null hypothesis (i.e., no difference in the pre- and post-ratings of B) to that under the alternative hypothesis (a one-tailed negative difference between the pre- and post-ratings of B). This analysis revealed that the observed difference between the pre- and post-ratings of B (Δ = -4.25) was 2.85 times more likely under the alternative hypothesis than the null hypothesis. Using Raftery (1995) and Jeffrey'(1961) guidelines for interpreting Bayes Factors, a Bayes Factor of 2.85 represents, respectively, weak and anecdotal evidence in support of the alternative hypothesis. Thus, the marginal difference between the pre- and post-ratings of B observed above does not represent a significant difference between the pre- and post-ratings of object B.

Given that some previous studies used the difference in the proportion of B choices in the BB and IS condition as evidence of BB reasoning (e.g., Sobel et al., 2004), the final analysis examined whether participants' post-ratings of B in the IS condition differed from their post-ratings of B in the BB condition. In particular, we examined whether participants' post-ratings of B in the IS condition were higher than their post-ratings of B in the BB condition. Consistent with previous blicket studies, this model revealed a significant main effect for Causal Rating, Type III *F*(1,59) = 237.44, *p* < .001, which indicated that participants in the IS condition provided higher post-ratings of B (β = 42.25, *M* = 42.25, Bootstrapped 95% CI[37.79, 46.71]) than did participants in the BB condition of B (Δβ = 48.07, *M* = 90.32, Bootstrapped 95% CI[41.62, 54.51]). This is perhaps not surprising because A was shown explicitly not to make the detector go in the IS condition. This means that participants should have judged B to be the cause. In contrast, A was shown to make the machine go in the BB condition, which should lead participants to judge A to be the cause. However, as we mentioned in the Introduction, this comparison may not be the most appropriate one to make to assess BB reasoning. Indeed, when the appropriate comparison was undertaken by comparing the pre- and post-ratings of B in the BB condition, this analysis demonstrated that the difference in the pre- and post-ratings of B in the BB condition was non-significant.

**Discussion**

The results from this experiment indicated that participants’ post-ratings of objects A and B differed reliably from their pre-ratings of both objects in all conditions except for the BB test trial, where the difference between the pre- and post-rating of object B was shown to be non-significant. This BB result is interesting because it suggests that adults may not have engaged in BB reasoning. However, as will be discussed next, this interpretation should be interpreted cautiously because this lack of a drop may well be interpreted as evidence of BB reasoning on some accounts but not others.

**Multiple interpretations of BB reasoning**

Despite the fact that adults did not appear to engage in BB reasoning when asked to reason about two objects it is worth noting a potential criticism of Experiments 1, 2, and 3. To this point, we have argued that, at least on some accounts, BB represents an absolute drop in the rating of B between the pre- and post-rating phases and subsequently interpreted the lack of an observed drop in the pre- and post-ratings of B across Experiment 1 to mean that adults did not engage in BB reasoning. However, if BB refers not to an absolute drop in the rating of B but instead to a return to baseline between the pre- and post-rating phases then the finding that adults’ pre- and post-ratings of B did not differ in Experiments 1-3 are consistent with predictions from a Bayesian inference account or certain associative accounts. For example, it may be the case that participants who engage in BB reasoning do so by blocking B between a mid- and post-rating phase rather than between the pre- and post-rating as we have assumed. Indeed, this conclusion is supported by Griffiths et al.'s (2011) finding that B’s causal rating was much more likely to undergo a drop between the mid-rating and post-rating phases than between the pre- and post-rating phases, where the two ratings were shown not to differ. This conclusion is also consistent with the predictions of the Bayesian model reported in Griffith et al.’s (2011) study: post-ratings of B were predicted to be lower than mid-ratings of it but no different than the pre-ratings of it. However, note that this conclusion must be interpreted cautiously in the context of the Experiment 1 because it is unclear whether such a drop would have occurred given that it was not directly assessed in the experiment.

Thus, because it is unclear what causal ratings adults should assign to B at each of the three time points if adults use an associative or Bayesian mechanism to reason about causal events, we first built a Bayesian model—based on that presented in Griffiths et al. (2011)—and two associative models; that is, a model that used the Rescorla-Wagner associative learning rule and one that used the modified Rescorla-Wagner associative learning rule (discussed below). The rationale for building these models was threefold. First, our aim was to assess what ratings participants should assign to object B at each of the three time points if they reasoned according to either of the three models. Second, we sought to examine whether, and to what extent, the three models differed in whether they predicted blocking (i.e., a drop in the rating of B) across the three rating time points. Third, we wanted to clarify the results from Experiment 1, which, at present, may be consistent with a Bayesian or an associative account.

**Computational Models and Predictions**

To clarify the results from Experiment 1 and to determine on what basis adults reasoned about the causal events presented here, we built a Bayesian model and two associative models. The two associative models we focused on were the Rescorla-Wagner model (henceforth, the RW model; Rescorla & Wagner, 1972) and the modified Rescorla-Wagner model (Van Hamme & Wasserman, 1994). We focused on the RW model because this model was among the earliest associative models of causal learning that, despite its successes, failed to account for BB. We also focused on this model because it has been argued by some that this model is unable to account for adults’ performance in a BB task and that Bayesian inference must underpin adults’ performance on this task (e.g., Sobel et al., 2004; Griffiths et al., 2011). However, as will be shown next, this model makes pre- and post-rating predictions for object B that were confirmed in Experiment 1 and that are partially in concert with those of a Bayesian model for two out of three rating phases.

The reason we focus on the modified RW model and a Bayesian model is because, unlike the RW model, both models conceptualize BB reasoning differently. Indeed, although both models interpret BB as a drop of some kind between different rating phases, the modified RW model interprets BB as an absolute drop in the rating of B between the pre- and post-rating phase, whereas the Bayesian model interprets BB as a return to baseline between these two rating phases. Lastly, we focus on all three models because they generate novel predictions that will be tested and verified in subsequent behavioral experiments. It is important to note that these models yield different qualitative predictions against which we compared the performance of adults across the four experiments. It should also be noted that although a more exhaustive approach could have been implemented in which we compared the all existing associative, rational-parameter models (e.g., the Power PC model), and Bayesian models, we focused on only these models because much of the recent debate in the literature tends to center on whether, and to what extent, these are viable models of causal cognition (though for a more extensive discussion on this topic and model comparison approach can be found in Sobel et al., 2004 and Griffiths et al., 2011).

Finally, it is worth mentioning that the models and simulations discussed below are intended to model Experiment 4 given that participants in this experiment, unlike those in Experiments 1 to 3, were asked to provide three causal ratings. By providing three causal ratings, it is possible to determine whether, and to what extent, adults distinguish the causal events based on Bayesian learning or associative learning.

**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. More 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 a given hypothesis generated that causal data. Figure 4 shows the possible hypotheses for Experiment 2. Following observations of data, *d*, the learner 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*. The denominator serves as the normalizing term, which normalizes the posterior probabilities 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 4) and is consistent with the observed data; otherwise, they are set to 0. Table 1 below displays the likelihoods for each hypothesis in Experiment 2.

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

Insert Figure 4 about here

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

The first step in defining a model of the task presented in Experiment 1 was to specify the hypothesis space H and the hypotheses h that comprise that space. This step is necessary if one seeks to use Bayes' rule to determine the hypothesis with the largest posterior probability. Because Experiment 2 uses two objects (i.e., two candidate causes), the hypothesis space consists of four hypotheses. The specific parameterization of each hypothesis in the space is specified by the activation law, which, for Experiment 2, 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. If we assume that the probability that a particular object is a blicket is independent of the probability that other objects are blickets, then the prior probabilities for hypotheses 1-4, respectively, are *p*2, *p*(1-*p*), *p*(1-*p*), and (1-*p*)2. These prior probabilities can then be used to compute posterior probabilities for each hypothesis when new data is observed according to Bayes' rule (see Table 1 for the prior and posterior probabilities of each hypothesis at different points in the BB condition).

To determine the prior probability that object A or B is a blicket in Experiment 4, a first requirement is to determine the likelihood, *p*(*d*|*h*), that object A or B is a blicket across each of the four hypotheses. In particular, we must determine whether a causal link exists between object A or B and the effect. 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 is a blicket by taking the product of the likelihood that object A (or B) 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 equals 1 if a causal link between *B* and *E* existsfor a specific hypothesis *h*; otherwise, equals 0.

The qualitative predictions of this model for how participants should rate A and B in the BB condition in Experiment 1 are shown in Table 1. It should be clear from this table that the model predicts that participants should provide higher causal mid-ratings of A and B than pre-ratings of both objects. The reason the model makes this prediction is because participants observe that together objects A and B make the detector activate when they are placed on its surface. This model further predicts that participants' post-ratings of A should be at or near ceiling and that participants should provide higher mid- and post-ratings of A than pre-ratings of it. Although the model predicts that participants should provide higher mid-ratings of B than pre-ratings of it, the pre- and post-ratings of B are not predicted to differ. Despite the fact that we do not explicitly manipulate the prior probability of being a blicket in any of the reported experiments, this model predicts that the increase in the rating of A and B between the pre- and mid-rating phases should become smaller as the prior probability of being a blicket increases.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 2 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 1. This table displays the predictions of the Bayesian model for how adults should rate objects A and B 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 qualitative predictions of this model for Experiments 3 and 4 are shown in Tables 2 and 3 below, respectively. Note that these experiments were similar to Experiment 2 except that adults were asked to reason about 3 (Experiment 3) and 4 (Experiment 4) objects.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 3 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 2. This table displays the predictions of the Bayesian model for how adults should rate objects A, B, and C 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 of the 8 graphs before the BB event, after the AB+ event, and then after the A+ event. The prior probabilities for all three 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 4 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 3. This table displays the predictions of the Bayesian model for how adults should rate objects A, B, and C 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 of the 8 graphs before the BB event, after the AB+ event, and then after the A+ event. The prior probabilities for all three 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.

Because participants in Experiment 5 were asked to rate four test events that included the A+, A-, B+, and B- events at each of three time points, the predictions for this experiment were similar, but not identical, to those for Experiment 2. In this experiment, adults were shown events in which one or two objects (A or B) caused a sun to rise from within a centrally located box when they contacted it in the four conditions. Subjects were then asked to rate how likely each of four test events were to occur on a scale between 0 (certainly unlikely) and 100 (certainly likely). These test events included an A+ test event (in which individually A caused the sun to emerge), an A- test event (in which individually A failed to cause the sun to emerge), a B+ test event (in which individually B caused the sun to emerge), and a B- test event (in which individually B failed to cause the sun to emerge) (see Experiment 5 for full procedural details). The model’s predictions for this experiment were thus that participants should provide higher mid-ratings of the A+ and B+ test events than their pre-ratings of the same events, whereas their mid-ratings of the A- and B- events were predicted to be lower than their pre-ratings of the same events. In addition, the model predicts that participants' post-ratings of the A+ and A- test events should be higher and lower, respectively, than their pre-ratings of it. In contrast, participants should provide lower and higher post-ratings of the B+ and B- test events, respectively, than mid-ratings of the same two test events. These predictions from this simple Bayesian model were tested in Experiment 4.  
These qualitative predictions are shown below in Table 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPERIMENT 5 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** | | | |
| A+ | *p* | 1/(2-*p*) | 1 |
| B+ | *p* | 1/(2-*p*) | *p* |
| A- | 1-*p* | 1-(1/(2-*p*)) | 0 |
| B- | 1-*p* | 1-(1/(2-*p*)) | 1-*p* |

Table 4. This table displays the predictions of the Bayesian model for how adults should rate the A+, A-, B+, and B- test events in Experiment 5 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 the four test events 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.

**Associative models**

*Rescorla-Wagner model*. In contrast to the Bayesian model that is presented above that uses Bayes’ rule to update the posterior probabilities of each of the candidate causal graphs, the Rescorla-Wagner model (Rescorla & Wagner, 1971) uses error correction to produce causal ratings of the test events. 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*, 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, we simulated Experiments 2-5 to examine what predictions the RW model generated and to determine whether those predictions differed from those of the Bayesian model for these experiments. Although we set the weight for A and B each arbitrarily to equal .5, it is worth mentioning that the predictions discussed below held to the extent that neither salience parameters nor the weights were set to 0. Furthermore, given that the RW model does not inherently espouse a parameter that accounts for base rates (i.e., prior probabilities), we set the weight for A and B each to equal.5 to reflect that fact that, averaged across both objects and all four conditions, participants’ pre-ratings of A and B (*M* = 50.04) did not differ reliably from 50%, *t*(479) = 0.05, *p* = 0.48 (one-tailed); that is, participants seemed implicitly to assume uncertainty about the causal status of the objects prior to the presentation of any of the condition events. Finally, the model received 20 AB+ training trials and 20 A+ training trials. Following these trials, A and B were presented alone and the model’s prediction about the causal status of both objects was assessed. Note 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 2-5. This approach is also consistent with that taken in Griffith et al. (2011).

|  |  |  |  |
| --- | --- | --- | --- |
| 1. **EXPERIMENT 2 RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 1 |
| B | *p* | *p* | *p* |
| 1. **EXPERIMENT 3 RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | *p* |
| C | *p* | *p* | *p* |
| 1. **EXPERIMENT 4 RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | *p* |
| C | *p* | *p* | *p* |
| D | *p* | *p* | *p* |
| 1. **EXPERIMENT 5 RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A+ | *p* | *p* | 1 |
| A- | 1-*p* | 1-*p* | 0 |
| B+ | *p* | *p* | *p* |
| B- | 1-*p* | 1-*p* | 1-*p* |

Table 5A-D. This table displays the Rescorla-Wagner model’s predictions of the causal ratings for Experiments 2 (A), 3 (B), 4 (C), and 5 (D) before the BB event, after the AB+ event, and then after the A+ event. The ratings in each cell can be thought to represent confidence that a particular object is a cause..

As can be seen from the Table 5 above, this model predicts that the ratings of object A and B in Experiment 2 (Table 5A), A, B and C in Experiment 3 (Table 5B), A, B, C, D in Experiment 4 (Table 5C), and the A+, A-, B+, and B- test events in Experiment 5 (Table 5D) should not differ between the mid-rating and post-rating phases. In addition, the model predicts that the ratings of the B, C, and D objects as well as the ratings of the B+ and B- test events should not change across the three rating phases. However, the model predicts increases in the ratings of object A and the A+ test event and decreases in the rating of the A- event between the mid- and post-rating phases. The reason the model predicts that the causal ratings of the causally redundant objects and events (i.e., objects B, C, D, and the B+ and B- test events) should remain unchanged across the three rating phases of the BB condition is because these objects were never presented in isolation and, as such, the salience parameters associated with each causally redundant object and test events equal 0; that is, the weights associated with these objects and test events are not updated for this trial. Indeed, this model requires that cues be present—which none of the causally redundant objects or test events are during the A+ phase of the BB event—in order for the associative weight between them and the outcome to be modified.

*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 terms of how the salience parameter, α, is treated. Whereas α is set to 0 for absent cues in the traditional Rescorla-Wagner 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 (Table 6).

|  |  |  |  |
| --- | --- | --- | --- |
| 1. **EXPERIMENT 2 MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
|  | Before AB+ | After AB+ | After A+ |
| A | *p* | *p* | 1 |
| B | *p* | *p* | 0 |
| 1. **EXPERIMENT 3 MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | 0 |
| C | *p* | *p* | 0 |
| 1. **EXPERIMENT 4 MODIFIED RESCORLA-WAGNER MODEL PREDICITONS** | | | |
| A | *p* | *p* | 1 |
| B | *p* | *p* | 0 |
| C | *p* | *p* | 0 |
| D | *p* | *p* | 0 |
| 1. **EXPERIMENT 5 MODIFIED RESCORLA-WAGNER MODEL PREDICTIONS** | | | |
| A+ | *p* | *p* | 1 |
| A- | 1-*p* | 1-*p* | 0 |
| B+ | *p* | *p* | 0 |
| B- | 1-*p* | 1-*p* | 1 |

Table 6A-D. 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 confidence that a particular object is a cause. Thus, a rating of 0 indicates absolute certainty that a particular object is not a cause, a 50% indicates uncertainty about a particular object’s causal status, and a rating of 100% indicates absolute certainty about the object’s causal status.

As can be seen in Table 6 above, the modified RW model, like the traditional model, predicts that ratings of the four objects and test events should remain unchanged between the pre- and mid-rating phases. The reason for this prediction is identical to that listed for the traditional RW model. That is, because ΣVA+B equals λ, the associative weights for A and B should remain unchanged between the pre- and post-rating phases. However, unlike the Bayesian or the RW models, the modified RW model predicts that the ratings of the causally redundant objects in Experiments 2-4. In Experiment 5, the model predicts that the ratings of the B+ and B- test events should undergo a drop and increase, respectively., whereas and test events should drop between the mid- and post-rating phases. As should be clear from the foregoing discussion, all three models make different qualitative predictions about how participants should rate the causally redundant cuesacross the pre-, mid-, and post-rating phases across the four experiments.

**Discussion**

The goal of Experiment 2 was to test the predictions of the Bayesian model and the traditional and modified RW models for two objects. Because Experiment 2 was identical to Experiment 1 in all respects except that participants were asked to provide three sets of ratings, it is possible to use the results of Experiment 2 to clarify those in Experiment 1; that is, to clarify on what basis adults’ processed the BB event. In line with the predictions outlined above, we predicted that if participants engaged in BB reasoning vis-à-vis the predictions of the Bayesian model, then (a) their mid-ratings of object B would be higher than their pre-ratings of it, (b) their mid-ratings of B would be higher than their post-ratings of it, and (C) their pre- and post-ratings of B would not differ. However, if participants used a rule similar to the RW rule to reason about causal events, it was predicted that their pre-, mid-, and post-ratings of object B should not differ. Finally, if adults used the modified Rescorla-Wagner learning rule to reason about causal events, their ratings of the test events should be similar to those that are predicted by the traditional RW model with one key exception: participants should provide lower post-ratings of object B than either their pre- or mid-ratings of it. In addition, participants’ pre- and mid-ratings of the B+ (and B-) test event were not expected to differ. Thus, unlike the Rescorla-Wagner model but similar to the Bayesian model, the modified model predicts that participants should block B following the A+ training event but not before. We tested these predictions below in Experiment 4.

**Experiment 2**

Experiment 2 was identical to the first experiment in all respects except that participants were asked to provide pre-, mid-, and post-ratings of objects A and B. This manipulation enabled us to assess how participants’ ratings of the two objects changed across the three rating phases and to determine on what basis adults reasoned about the events. That is, to the extent that the pre- and post-ratings of adults in Experiment 2 replicate those in Experiment 1, it was possible to determine whether adults overall pattern of ratings for objects A and B were consistent with the predictions of the Bayesian model or with either of the two associative models and thus to determine whether adults used Bayesian inference or associative processing to evaluate the test events.

**Experiment 5**

The aims of Experiment 4 were threefold. First, we examined whether, and to what extent, adults engaged in BB reasoning across a pre-, mid-, and post-rating phase. Second, we examined whether, and to what extent, adults engaged in BB reasoning in a novel task that used video sequences as opposed to the physical blicket detector and that required participants to rate four different test events. Third, Experiment 4 was designed to test the predictions of the models described above to determine on what basis adults reason about causal events and to clarify the results from Experiments 1 to 3.

**Method**

**Participants.** Sixty college students were recruited for Experiment 4 and were given course credit for their participation. There were approximately equal numbers of males and females in the final sample.

**Stimuli and Design.** The training and test events were computer-animated events that were created on the personal computer (PC) with Macromedia Director 8.0. The movies were presented on a computer-generated orange stage with a light purple background. The four training event movies each consisted of a red and green circle, each approximately 86 pixels in circumference, and square box, approximately 173 pixels wide by 103 pixels in height, with a yellow star drawn in the middle of the box. The purpose of the star in the middle of the square was to make the object more perceptually salient and thereby to increase participants' interest in the box.

**Procedure.** Participants were tested in a quiet room on the campus of Carnegie Mellon University. At the beginning of the experiment, participants were told first that they would be shown four training event sequences; these events corresponded to the BB, IS, 1C, and 2C conditions. Note that participants were merely told that they would be shown four test events and thus did not receive information about the kinds of events that they were to be shown. Participants were then instructed to rate at three different points—before a particular training event, midway through the training event, and following the full training event—how likely each of four test events were to occur following the presentation of a particular training event. In particular, participants were then asked to rate how likely each of four test events were to occur on a scale between 0 (certainly unlikely) and 100 (certainly likely). Participants were encouraged to use the entire range.

The training and test stimuli were computer-animated and were created on the personal computer (PC) with Macromedia Director 8.0. The movies were presented on a computer-generated orange stage with a light purple background. The four training movies each consisted of a red and blue circle, each approximately 86 pixels in circumference, and square box, approximately 173 pixels wide by 103 pixels in height, with a yellow star drawn in the middle of the box. The purpose of the star in the middle of the square was to make the object more perceptually salient and thereby to increase participants' interest in the box.

In the BB training event, a red and blue circle entered the stage from the right and left (counterbalanced), moved horizontally across the stage (at a rate of approximately 30 fps) for approximately 172 pixels, at which point the red and green circles became adjoined with the square that was located in the middle of the screen. When both objects were adjoined with the box, a sun popped out of the box and ascended upwards for 297 pixels, at which point the sun engaged in a bouncing motion for 1 s before stopping while suspended above the box. The sun was approximately 192 pixels wide by 123 pixels in height. Participants were then shown that one of the two objects—which represented object A (counterbalanced)—caused the sun to emerge when it made contact with the centrally located square. For notation purposes, this event was labeled the "AB+; A+" event, where A and B denote the objects and the "+" refers to the presence of the effect (i.e., the sun). The IS training event was identical to the BB training event except that A failed to make the sun emerge when alone it contacted the box. This event was labeled the "AB+; A-" event. In the 1C training event, object A made the sun emerge by itself, whereas object B did not. Both objects then made the sun emerge twice when they made contact with the box from the right and left both times. This event was labeled the "A+ B-; AB+" event. Finally, in the 2C training event object A made the sun emerge by itself three times, whereas object B failed to make the sun emerge the first time but then made the sun emerge when it made contact with the box on two additional occasions. This event was labeled the "A+++; B-; B++" event. Participants were shown each training event three times to ensure sufficient encoding of the events.

Following each training event, participants were shown four test events: A+, A-, B+, B-. In the A+ event, A made the sun emerge when alone it contacted the box. The B+ test event was similar to the A+ test event, except that object B, but not object A, caused the sun to appear when it became adjoined with the box. Finally, the A- and B- events were similar to the A+ event, except that when objects A (in the A- event) and B (in the B- event) became adjoined with the box, the sun did not appear. Finally, The training events and test events each lasted 6 s.

**Results**

**Control condition: 1C.** The first analysis examined whether adults engaged in 1C and 2C reasoning. Similar to the previous experiments reported here, these two conditions served as the control conditions. We first examined whether adults' ratings of the A+ event differed across the pre-, mid-, and post-rating phases. Note that if adults engaged in 1C reasoning, they should provide higher mid- and post-ratings of the A+ event than pre-ratings of it and higher post-ratings of the same event than mid-ratings of it. A mixed-effects model was thus fit to determine whether participants' ratings of the A+ event changed across the three rating phases. This analysis revealed a significant main effect of Causal Rating, Type III *F*(2,59) = 65.03, *p* < .001. Planned comparisons using non-parametric permutation tests and bootstrapping revealed that participants' provided higher mid-ratings of A+ (*M* = 90.33, Bootstrapped 95% CI[85.53, 95.13]) than pre-ratings of it (*M* = 61.58, Bootstrapped 95% CI[56.52, 66.64]), *p* < .001, although participants' post-ratings of A+ (*M* = 92.08, Bootstrapped 95% CI[88.33, 95.84]) did not differ significantly from their mid-ratings of it, *p* = .43.

We next entered participants' ratings of the A- event into a mixed-effects model, which revealed a significant main effect of Causal Rating, Type III *F*(2, 59) = 81.65, *p* < .0001. Note that adults' ratings of the A- event should be the inverse of their ratings of the A+ event and, in particular, should show a monotonic decrease across the three rating phases. Planned comparisons corroborated this conclusion and revealed that participants provided lower mid-ratings of the A- event (*M* = 16.67, Bootstrapped 95% CI[9.93, 23.4]) than pre-ratings of it (*M* = 48.58, Bootstrapped 95% CI[46.11, 51.06]), *p* < .0001. Likewise, participants provided lower post-ratings of the A- event (*M* = 12.08, Bootstrapped 95% CI[6.83, 17.33]) than mid-ratings of it, *p* < .001.

Participants' ratings of the B+ event were entered into a mixed-effects model, which revealed a significant main effect of Causal Rating, Type III *F*(2, 59) = 157.97, *p* < .0001. Similar to the ratings of the A- event across the three rating phases, adults' ratings of the B+ event should show a monotonic decrease across the three rating phases. Planned comparisons confirmed this prediction and revealed that participants provided lower mid-ratings of the B+ event (*M* = 15.07, Bootstrapped 95% CI[9.19, 20.95]) than pre-ratings of it (*M* = 62, Bootstrapped 95% CI[56.9, 67.09]), *p* < .0001, although their mid-ratings of that event did not differ from their post-ratings of it (*M* = 14.5, Bootstrapped 95% CI[9.31, 19.69]), *p* = 1.

Finally, we entered participants' ratings of the B- event into a mixed-effects model. Note that adults' causal ratings of this event, similar to the A+ event, should show a monotonic increase across the pre-, mid-, and post-rating phases. This was confirmed by a significant main effect of Causal Rating, Type III *F*(2, 59) = 113.74, *p* < .0001. Planned comparisons revealed that participants provided higher mid-ratings of the B- event (*M* = 84.28, Bootstrapped 95% CI[77.67, 90.89]) than pre-ratings of it (*M* = 48.08, Bootstrapped 95% CI[45.88, 50.28]), *p* < .0001. However, participants' post-ratings of the B- event (*M* = 84.07, Bootstrapped 95% CI[77.79, 90.34]) did not differ from their mid-ratings of that event, *p* = 1.

**Control condition: 2C.** The second analysis examined whether, and to what extent, adults' ratings of the four test events changed across the pre-, mid-, and post-rating phases in the 2C condition. We first entered participants' ratings of the A+ into a mixed-effects model. This analysis revealed a significant main effect of Causal Rating, Type III F(2, 59) = 120.05, p < .0001. Because A produces the effect each time it is presented in the A+, it was predicted that if participants engaged in 2C reasoning their causal ratings of this event would undergo a monotonic increase across the three rating phases. Planned comparisons confirmed this prediction and revealed that participants provided higher mid-ratings of the A+ event (*M* = 90.67, Bootstrapped 95% CI[85.21, 96.12]) than pre-ratings of it (*M* = 53.92, Bootstrapped 95% CI[50.28, 57.45]), *p* < .0001. However, participants' post-ratings of the A+ event (*M* = 93.75, Bootstrapped 95% CI[89.89, 97.61]) did not differ from their mid-ratings of it, *p* = .36.

We next entered adults' ratings of A- event into a mixed-effects model to determine whether those ratings changed across the three rating phases. This analysis revealed a significant main effect of Causal Rating, Type III *F*(2, 59) = 47.94, *p* < .01. In contrast to participants’ ratings of the A+ events, participants’ ratings of the A- event should show a monotonic decrease across the three rating phases. Planned comparisons revealed that participants provided lower mid-ratings (*M* = 16.83, Bootstrapped 95% CI[9.33, 24.33]) than pre-ratings of it (*M* = 49, Bootstrapped 95% CI[43.23, 54.77]), *p* < .0001, and lower post-ratings of the same event (*M* = 12.67, Bootstrapped 95% CI[5.94, 19.39]) than mid-ratings, *p* < .0001.

In addition, we entered adults' ratings of the B+ event to determine how it changed across the three rating phases. This model revealed a significant main effect of Causal Rating, Type III *F*(2, 59) = 30.39, *p* < .0001. Because participants were asked to provide two ratings of B (and A) before B’s causal status was known, it was predicted that participants’ mid-ratings of the B+ event should be lower than their pre-ratings. This is because A would have been shown to produce the effect on three separate instances which, in turn, may cause participants erroneously to discount B. However, participants’ ratings of that event should undergo an increase between the mid- and post-rating phases once they observe that B causes the sun to rise. This prediction was confirmed by planned comparisons, which revealed that participants provided higher pre-ratings of the B+ event (*M* = 53.58, Bootstrapped 95% CI[49.92, 57.25]) than mid-ratings (*M* = 47.83, Bootstrapped 95% CI[43.04, 52.63]), *p* < .0001, but higher post-ratings (*M* = 66.85, Bootstrapped 95% CI[64.39, 69.31]) than either pre-, *p* < .0001, or mid-ratings, *p* < .0001.

Finally, we entered adults' ratings of the B- event into a mixed-effects model to examine how these ratings changed across the three rating phases. This model did not yield a significant main effect of Causal Rating, Type III *F*(2, 59) = 6.76, *p* < 0.005. Despite the fact that main effect was not significant, we nonetheless conducted planned comparisons. These analyses revealed that participants' pre- (*M* = 48.17, Bootstrapped 95% CI[42.68, 53.65]) and mid-ratings (*M* = 45.97, Bootstrapped 95% CI[39.88, 52.05]) of the B- event did not differ, *p* = 1, whereas participants provided lower post-ratings (*M* = 35.93, Bootstrapped 95% CI[31.92, 39.95]) of the B- event than mid-ratings, *p* < .0001, or pre-ratings, *p* < .0001.

In general, these results represent conceptual replications of Sobel et al.’s (2004) study and suggest that adults engaged in 1C and 2C reasoning in this task that used video sequences.

**IS condition.** The first analysis fit a mixed-effects model to participants’ ratings of the A+ event across the three rating phases to determine whether and to what extent participants' ratings of this event differed across the three rating phases. This model demonstrated a main-effect for Causal Rating, which indicated that participants' causal ratings of the A+ event differed across the three rating phases, Type III *F*(2, 59) = 18.38, *p* < .0001. Follow-up planned comparisons revealed that participants provided marginally higher mid-ratings of A+ (*M* = 54.15, Bootstrapped 95% CI[49.38, 58.92]) than pre-ratings (*M* = 57.17, Bootstrapped 95% CI[42.14, 72.19]), p = 08. In contrast, participants provided lower post-ratings of A+ (*M* = 18.95, Bootstrapped 95% CI[12.95, 24.95]) than mid-ratings, *p* < .00001, or pre-ratings, *p* < .0001, of the same event. That participants' post-ratings of A+ were lower than their pre- or mid-ratings is consistent with the evidence in the training event because A was shown not to make the sun rise in the IS condition. Note that this finding predicts that the A- should be considered more likely as the event unfolds and thus the distribution of the A- responses should be the inverse of the A+ responses across the three rating phases.

To explore this possibility, we fit a separate mixed-effects model to determine whether participants' ratings of the A- event differed across the three ratings phases and, importantly, whether these ratings were the inverse of those of the A+ event. This analysis revealed a significant main effect of Causal Rating, Type III *F*(2, 59) = 69.62, *p* < .0001. Although participants' pre-ratings of A- event (*M* = 48.92, Bootstrapped 95% CI[43.62, 54.22]) differed marginally from their mid-ratings of it (*M* = 43.07, Bootstrapped 95% CI[37.45, 48.68]), *p* = 0.67, participants did provide significantly higher post-ratings of the A- event (*M* = 83.23, Bootstrapped 95% CI[76.63, 89.84]) than either the pre-ratings or mid-ratings of it, both *p*'s < .0001.

We next fit a mixed-effects model to examine whether participants' ratings of the B+ event differed across the pre-, mid-, and post-rating phases. This model yielded a significant main effect of Causal Rating, Type III *F*(2, 59) = 52.83, *p* < .0001. Planned comparisons revealed that participants provided significantly higher mid-ratings (*M* = 57.77, Bootstrapped 95% CI[53.31,62.63]) than pre-ratings (*M* = 49.67, Bootstrapped 95% CI[46.8,52.53]) of the B+ event, *p* < .01, and significantly higher post-ratings (*M* = 80.17, Bootstrapped 95% CI[73.64,86.69]) than mid- or mid-ratings of it, both *p*’s < .0001. That participants' ratings of B+ increased monotonically across the three rating phases is consistent with the evidence because it suggests that participants reasoned that if A did not cause the sun to emerge B must be the cause and that participants perceived the B+ event as more likely as the IS condition unfolded.

We fit a final mixed-effects model to examine whether participants' causal ratings of the B- event changed across the three rating phases and to assess whether ratings of the B- event were the inverse of those of the B+ event. This model yielded a significant effect of Causal Rating, Type III *F*(2,59) = 45.36, *p* < .001. Planned comparisons revealed that although the pre- (*M* = 48.83, Bootstrapped 95% CI[43.51, 54.15]) and mid-ratings (*M* = 46.02, Bootstrapped 95% CI[40.11, 51.92]) of the B- event did not differ, p = 0.24, participants provided lower post-ratings (*M* = 19.33, Bootstrapped 95% CI[12.28, 26.39]) of the same event than either pre- or post-ratings, both *p*'s < .0001. Figure 4 below shows the change in the ratings of the A+, A-, B+, and B- test events across the pre-, mid-, and post-rating phases.

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

Insert Figure 4 about here

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

**BB condition.** To explore whether participants' ratings of the A+ event differed across the three rating phases, we again fit a mixed-effects model. This model revealed a significant main-effect of Causal Rating, Type III *F*(2, 59) = 48.56, *p* = .0001, whereby participants' pre-ratings of the A+ event (*M* = 65.25, Bootstrapped 95% CI[59.56, 70.94]) were significantly higher than their mid-ratings of it (*M* = 58.15, Bootstrapped 95% CI[54.44, 61.86]), *p* < .05. Participants also provided higher post-ratings of the A+ event (*M* = 90.15, Bootstrapped 95% CI[86.15, 94.15]) than pre- and mid-ratings of it, both *p*’s < .0001. This finding suggests that participants perceived the A+ event as more likely as the BB event unfolded, which is consistent with the observation that A caused the sun to rise throughout the BB event.

Given that A was shown to produce the effect in the BB condition, this predicts that participants should have reasoned that the A- event was less likely as the BB event unfolded. To examine this prediction, we next fit a mixed-effects model using participants' pre-, mid-, and post-ratings of the A- event. This analysis revealed a significant main effect of Causal Rating, Type III *F*(2, 59) = 92.41, *p* < .0001. Planned comparisons revealed that participants provided lower mid-ratings of the A- event (*M* = 43.56, Bootstrapped 95% CI[38.79, 48.34]) than pre-ratings of it (*M* = 48.92, Bootstrapped 95% CI[45.32, 52.51]), *p* < .05. Likewise, participants provided significantly lower post-ratings of the A- event (*M* = 15.42, Bootstrapped 95% CI[9.5, 21.33]) than either the pre-ratings or mid-ratings of it, both *p*’s < .0001.

The next two analyses were perhaps the most important for three reasons. First, they examined whether participants across the BB event provided higher mid-ratings of the B+ event than either pre- or post-ratings of it (which themselves should not differ). Second, the analyses examined whether adults provided lower mid-ratings of the B- event than either the pre- or post-ratings of it, and higher post-ratings of the B- event than either the pre- or-mid ratings of it. Evidence for both predictions—which are consistent with those of a Bayesian inference account—are required to support the contention that adults engaged in BB reasoning in the present task. Third and relatedly, the analyses have the potential to corroborate the predictions of one of the models presented above and thus have the potential to provide insight into the learning mechanism that underpins learning in adults.

The first analysis focused on participants' ratings of the B+ event across the pre-, mid-, and post-rating phases. Participants' ratings of the B+ event were similarly entered into a mixed-effects model, which revealed a significant main effect of Causal Rating, Type III *F*(2, 59) = 137.01, *p* < .0001. Planned comparisons revealed that participants provided lower mid-ratings of the B+ event (*M* = 56.82, Bootstrapped 95% CI[52.91, 60.72]) than pre-ratings of the same event (M = 63.25, Bootstrapped 95% CI[57.9, 68.59]), *p* < .05, and lower post-ratings of the same event (*M* = 46.63, Bootstrapped 95% CI[41.02, 52.24]) than either the pre- or mid-event, *p*'s < .001. To assess the qualitative fit of the three models with participants’ ratings of the B+ event across the three rating phases, we first compared participants’ pre- and post-ratings of the B+ event. This analysis revealed that participants provided lower post-ratings (*M* = 46.63, Bootstrapped 95% CI[41.02, 52.24]) of the B+ event than pre-ratings of it (*M* = 63.25, Bootstrapped 95% CI[57.9, 68.59]), *p* < .001. It will be recalled that the modified RW model predicts that participants should provide a lower post rating of the B+ event than pre-ratings of it, whereas both the traditional RW model and the Bayesian model predict that the pre-ratings and post-ratings of the B+ event should be equivalent. Thus, this results supports the predictions of the modified RW model and not those of the traditional RW model or the Bayesian model.

We next compared participants’ pre- ratings of the B+ event to their mid-ratings of it. Note that, unlike the first comparison, only the Bayesian model makes a unique prediction for how participants’ ratings of the B+ event should change between the pre- and mid-rating phases; that is, the Bayesian model predicts that participants should provide higher mid-ratings of the B+ event than pre-ratings of it following the observation that objects A and B, together, caused the sun to rise. In contrast, both the traditional RW model and the modified model predict that there should be no difference between participants’ pre- and mid-ratings of the B+ event. This analyses revealed that participants provided significantly lower mid-ratings of the B+ event (*M* = 56.82, Bootstrapped 95% CI[52.91, 60.72]) than pre-ratings (*M* = 63.25, Bootstrapped 95% CI[57.9, 68.59]), *p* < .05. To determine whether this difference was meaningful, we estimated a Bayes factor. The Bayes Factor for this difference indicated that the data were 6.13 times more likely to result from a real difference than from no difference (i.e., null effect). This represents positive (e.g., Raftery, 1995) or substantial (e.g., Jeffrey, 1961) evidence that the observed difference between the pre- and mid-ratings of the B+ event was meaningful. Importantly, this result neither supports the predictions of the modified RW model or the traditional RW model—which both predict equivalent pre- and mid-ratings of the B+ event—nor those of the Bayesian model—which predicts that participants should provide higher mid-ratings of the B+ event than pre-ratings of it.

Finally, we compared participants’ mid-ratings of the B+ event to their post-ratings of it. Here, only the traditional RW model makes a unique prediction; that is, unlike the modified RW model or the Bayesian model which predict that participants should provide lower post-ratings of the B+ event than mid-ratings of it, the traditional RW model predicts that participants’ mid- and post-ratings of the B+ event should be equivalent. This analysis revealed that participants provided lower post-ratings of this event (*M* = 46.63, Bootstrapped 95% CI[41.02, 52.24]) than mid-ratings of it (*M* = 56.82, Bootstrapped 95% CI[52.91, 60.72]), *p < .*001. This result confirms the predictions of the Bayesian and modified RW models but not those of the traditional RW models. Together, these results suggest that participants used a version of the modified RW learning rule to reason about the B+ test event. This is because 2 out of the 3 predictions for this model were confirmed, whereas 0 and only 1 out of 3 predictions were confirmed, respectively, for the traditional RW model and the Bayesian model. However, this conclusion must be interpreted cautiously because the results did not support all three predictions of this model—hence why participants were said to have used a version of the modified RW model.

These results notwithstanding, evidence that participants used a Bayesian mechanism to reason about causal events requires that two conditions are met. First, participants must provide higher mid-ratings of the B+ event than either the pre- or post-ratings—which themselves may or may not differ—as the BB event unfolds. Second, participants must provide lower mid-ratings of the B- event than pre-ratings of it but higher post-ratings of the B- event than either the pre- or mid-ratings as the BB event unfolds. This means that support for either the first or the second condition, but not for both, would be insufficient to conclude that adults used a Bayesian mechanism to reason about causal events.

Thus, the next analysis examined whether participants provided lower mid-ratings of the B- event than pre-ratings of it but higher post-ratings of the B- event than either the pre- or mid-ratings of the same event as the BB condition unfolded. Unlike the previous analysis, this analysis did not reveal a significant main effect of Causal Rating, Type III *F*(2, 58) = 4.1, *p* < .05. We next conducted a series of planned analyses to explore this significant main effect. These analyses revealed that participants provided higher pre-ratings of the B- event (*M* = 48.25, Bootstrapped 95% CI[45.19, 51.3]) than mid-ratings of it (*M* = 41.9, Bootstrapped 95% CI[37.22, 46.58]), *p* = .01, and higher post-ratings (*M* = 51.5, Bootstrapped 95% CI[44.38, 58.62]) than mid-ratings, *p* = 0.015 . However, participants' post-ratings of the B- event (*M* = 51.5, Bootstrapped 95% CI[44.38, 58.62]) did not differ from their pre-ratings of it, p = .2.

Note that the predictions that each of the three models make for the B- event should be the inverse of those described above for the B+ event. In particular, the Bayesian model, unlike the traditional RW or modified RW models, predicts that there should be no difference in the pre- and post-rating of the B- event and that the ratings of these two events should be higher than the mid-ratings of the same event. Both the traditional RW model and the modified RW model predict that the pre- and mid-ratings of the B- event should differ. However, the traditional RW model, but not modified RW model, predicts further that post-ratings of the B- event should not differ from the pre- or mid-ratings of it. In contrast, the modified RW model predicts that participants should provide higher post-ratings of the B- event than either pre- or mid-ratings. Thus, the results for the B- event, unlike those for the B+ event, support the predictions of the Bayesian model.

Taken together, these results suggest that the evidence is mixed about whether, and to what extent, adults use Bayes rule or a modified version of the RW learning rule to reason about causal events. This is because participants’ pattern of responses to the B+ event across the three rating phases supported the predictions of the modified RW model, whereas participants’ pattern of responses to the B- event across the same three phases supported the predictions of the Bayesian model. Despite the fact that it is unclear whether participants used Bayes’ rule or a version of the modified RW rule to reason about causal events such as those presented here, it is clear that participants did not use the traditional RW learning rule to reason about causal events given that the predictions of this model for the B+ and B- events were not supported.

Given that previous research that used the difference in the proportion of children who chose B in the BB and IS condition to argue that children engaged in BB reasoning (e.g., Sobel et al., 2004), the final two analyses examined whether participants' post-ratings of the B+ and B- test events in the BB and IS condition. Despite the fact that we argued that this is the incorrect comparison, it is important to demonstrate that participants in our experiment also provided different ratings across these two conditions. This is because it would be difficult to interpret the present results in light of the previous blicket-detector findings if adults failed to provide different ratings of these events across conditions.

Thus, we first compared participants' post-ratings of the B+ event in the BB and IS conditions. This analysis revealed that participants provided lower post-ratings of the B+ event in the BB condition (*M* = 46.63, Bootstrapped 95% CI[41.02, 52.24]) than of the same event in the IS condition (*M* = 80.17, Bootstrapped 95% CI[73.64,86.69]), *p* < .0001. The second analysis revealed that participants provided higher post-ratings of the B- event in the BB condition (*M* = 51.5, Bootstrapped 95% CI[44.38, 58.62]) than of the same event in the IS condition (*M* = 19.33, Bootstrapped 95% CI[12.28, 26.39]), *p* < .0001. That participants' post-ratings of the B+ and B- events in the BB and IS conditions differed—which can be thought of as a conceptual replication of the findings with children that use the blicket detector—suggests that the IS, 1C, and 2C results discussed above represented real effects.

**Discussion**

Experiment 4 was designed to test the predictions of the Bayesian model and the two associative models. The results demonstrated that adults engaged in BB and provided partial support for the predictions of the Bayesian model and those of the modified RW model. In particular, participants provided lower mid-ratings than pre-ratings and lower post-ratings than mid-ratings. This pattern of responding was entirely consistent with the predictions of the modified RW model, which predicted a monotonous decrease in the causal ratings of the B+ event (and the inverse for the B- event). In contrast, participants provided lower mid-ratings of the B- event than either the pre- or post-ratings of it. This pattern of responding was entirely consistent with the predictions of the Bayesian model, which predicted a U-shaped pattern of responding to the B+ event and a reverse U-shaped pattern of responding to the B- event. These findings—which we discuss in detail in the General Discussion—are noteworthy for two reasons. First, these findings provided partial support for the Bayesian model and the modified RW model, which suggests that learners may use both Bayesian inference and associative processing to reason about causal events. This possibility is discussed in more detail in the General Discussion. Second, these finding suggests that, in contrast to previous discussions in the literature (e.g., Sobel et al., 2004), BB reasoning is not unequivocal evidence that humans use Bayesian inference. Rather, BB reasoning may represent a domain-general general phenomenon that is differently predicted by Bayesian and certain associative models (e.g., the modified RW model). This is because certain associative models such as the modified RW model and Bayesian models both predict BB, although both kinds of models predict that learners will engage in BB in qualitatively different ways: the modified RW model predicted a decrease in the rating of the B+ event between the mid- and post-rating phases but not between the pre- and mid-rating phases, whereas the Bayesian model predicted a U-shaped pattern of responding to the B+ event (and an inverse U-shaped pattern of responding to the B- event). It is important to note that the differential pattern of responding to the B+ and B- event cannot be attributed to a failure to replicate given that results from the 1C, 2C, and IS conceptually replicated previous research (e.g., Sobel et al., 2004). A potential explanation for this differential responding to the B+ and B- events and for how the results from Experiment 4 and those from Experiments 1 to 3 can be reconciled is discussed in the General Discussion.

**General Discussion**

The main goals of the experiments were threefold. First, the experiments were designed to examine whether adults engaged in backward-blocking reasoning using pre- and post-rating phases in Experiments 1 to 3 and pre-, mid-, and post-rating phases in Experiment 4. In particular, these experiments examined whether participants attributed less causality to causally redundant cues when it was revealed that another cue can produce the effect by itself. Second, they were devised to investigate whether and to what extent adults engaged in BB reasoning in the context of two, three, and four objects and whether such reasoning generalizes to a novel context that used animated video sequences. Third, these experiments were designed to determine whether a Bayesian mechanism or associative mechanisms underpin causal reasoning; that is, these experiments assessed whether a Bayesian inference mechanism or one of two associative mechanisms based on the learning mechanisms of traditional RW model and modified RW models explained the findings reported here. To our knowledge, this was the first study to assess BB reasoning directly by assessing the change in the rating of causally redundant cues across different phases, in the context of multiple objects and positive and negative events, and in events that use physical and computer animated objects.

These issues were addressed by asking adults to provide pre- and post-ratings of two, three, and four objects in Experiments 1 to 3 and pre-, mid-, and post-ratings of two objects in an experiment that used animated video sequences with two objects. Of particular interest was whether participants provided lower pre- and post-ratings of object B—the causally redundant cue—in Experiments 1 to 3 and whether participants' ratings of object B differed across the three rating phases in Experiment 4. We focused on object B because BB refers to the discounting of causally redundant cues. We also developed and tested the predictions of a Bayesian model, the traditional RW model, and the modified RW model to determine which model best accounted for adults' causal ratings of object B across three different rating phases in Experiment 4 and to clarify the results from Experiments 1 to 3. Experiments 1 to 3 revealed that adults did not provide lower post-ratings of object B than pre-ratings—regardless of whether they were asked to rate two, three, or four objects—and Experiment 4 revealed that adults engaged in BB reasoning. That participants’ responses to B confirmed the predictions of the Bayesian model and the modified RW model suggests that in all four experiments reported here that adults may use both Bayesian inference and associative processing when asked to reason about causal events and redundant causal cues.

These experiments are important because they are the first to address key but yet unanswered questions in the causal literature. Specifically, these experiments not only attempt to confirm the claim that humans can engage in BB reasoning but they address whether BB reasoning depends on the number of objects about which learners are asked to reason, whether these adults are asked to reason about physical objects or animated objects, and whether adults are asked to make inferences about positive or negative events. These experiments also extend previous research because they directly assess BB by comparing causal ratings of a redundant cue within, rather than between, a condition. Recall that this approach differs crucially from previous attempts in the literature that based the claim that humans engaged in BB 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 studies are also significant because they examine whether BB generalizes to novel but related contexts such as that in experiment 4 in which adults were asked to reason about animated causal objects. Perhaps most importantly, these experiments, and especially the computational models, are meaningful because they suggest that BB may be a domain-general phenomenon that is differentially predicted by Bayesian and certain associative models; that is, these experiments highlight that BB can be taken as evidence for a Bayesian account or an associative account. Note that this contrasts with previous accounts in which it is argued that BB reasoning is evidence that learners use Bayesian inference to reason about causal events. For example, in terms of the simple Bayesian model presented here, BB refers to a drop in the rating of the causally redundant cue between a mid-rating phase and a post-rating phase but no corresponding drop between a pre- and post-rating phase. In contrast, in terms of the modified RW model, BB refers to a drop in the rating of the redundant cue between the pre- and post-rating phases but not between the mid- and post-rating phases.

One question that remains unaddressed is whether, and to what extent, a Bayesian mechanism or an associative mechanism based on the modified RW model learning rule, or both, explained the present findings and provide the best account of causal reasoning more generally. One potential, albeit speculative, answer is that whether the adults in our experiments are said to have used Bayesian inference or associative processing depended on multiple factors. For example, although we argued that adults may have used both Bayesian inference and associative processing to reason about the potential causes in Experiments 1 to 3, it may have been the case that as the number of potential causes increased adults were less likely to use Bayesian inference. This may have to do with the fact that as the number of potential causes increases, so too does the size of the hypothesis space. Indeed, it is a known fact that the size of the hypothesis space is thought to increase exponentially as more potential causes are added (e.g., Gopnik & Wellman, 2012). Thus, it may have been the case that the adults in the present experiments used Bayesian inference to reason about a small number of causes because, in this situation, it is possible to enumerate the space of potential hypotheses; that is, it is possible—and indeed relatively trivial—to represent all possible hypothesis for two (size 4) or three (size 8) objects. However, as the number of potential causes increases—which others have labeled the "search problem" (e.g., Bonawitz, Denison, Gopnik, & Griffiths, 2014; Gopnik et al., 2004)—adults may be increasingly less likely to use Bayesian inference to reason about causal events due to information-processing limitations that prevent them from representing and enumerating all possible potential causal hypotheses. As such, it may have been the case in these experiments that adults were more likely to use associative processing to reason about four objects than they were to use it to reason about two or three objects but this possibility requires further testing.

This account implies that causal events may be processed along a causal gradient and that the mechanism that is deployed may critically depend on where one is along this causal gradient. For example, with a small number of causes, adults may first use Bayesian inference. When the number of potential causes increases slightly, adults may then be more likely to use a combination of both associative and Bayesian mechanisms, as may have been the case in the present experiments. However, as the number of potential causes increases still further, adults may shift to using associative processing to reason about causal events. This is an important issue that should be addressed in future studies that use a wider number of objects than that used here. It is worth mentioning that this account, although speculative, represents an important departure from previous accounts of causal learning and BB reasoning. Whereas it has been asserted in previous accounts that either a Bayesian mechanism (e.g., Gopnik et al., 2004) or an associative mechanism (e.g., Kloos & Sloutsky, 2013; Van Hamme & Wasserman, 1994) underpins causal processing, but not both, we leave open the possibility that both mechanisms may be operative during causal learning.

Finally, it is worth noting some potential criticisms of the present experiments. First, it is difficult to know definitively whether adults in Experiments 1 to 3 engaged in BB reasoning in a way that was consistent with the predictions of a Bayesian model or an associative model. This is because, unlike Experiment 4 in which we assessed BB across pre-, mid-, and post-rating phases, Experiments 1 to 3 only implemented pre- and post-rating phases. Thus, the observed lack of a difference between the causal ratings of the causally redundant cue in Experiments 1 to 3 could have resulted either from a Bayesian inference mechanism or an associative mechanism that used the traditional RW learning rule. The reason these findings do not provide support for the modified RW model is because this model predicts that participants' post-ratings of the redundant cue should be lower than their pre- and mid-ratings of it, which we did not observe in Experiments 1 to 3. Despite the fact that it is difficult to rule out this possibility, the fact that findings from Experiment 4 provided partial support for the predictions of the Bayesian model and the associative model and the fact that there is a rich body of literature that show that children (e.g., Sobel et al., 2004), adults (Griffiths et al., 2011; Lovibond et al., 2003; Shanks, 1985) and, to some extent, infants (e.g., Sobel & Kirkham, 2005) can engage in BB reasoning suggest that adults in Experiments 1 to 3 may have also engaged in BB reasoning. In fact, in the Discussion section following Experiment 3 we do not rule out this possibility. Nevertheless, future studies would benefit by implementing pre-, mid-, and post-rating phases to assess whether adults' BB performance supported a Bayesian mechanism or different associative mechanisms. Note that it is not sufficient to implement just a mid- and post-rating phase because, as discussed in the Computational Models and Predictions section, the Bayesian model and associative models made competing predictions about how adults should rate the causally redundant cue across these three rating phases.

Second, the experiments described here were designed to address an important limitation of previous BB 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 reasoning, these experiments were designed directly to examine BB reasoning. However, the experiments reported here do not directly address this issue in previous experiments 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 interesting 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 must preserve the logic of the experiments reported here and, more importantly, the logic of the BB event. The experiments reported here are nonetheless important because they provide insight into how adults reason about causal events, especially those used here, whether adults can engage in BB, and about the nature of the causal mechanism that underlies this ability and that can explain BB reasoning.

Finally, although we chose to focus on a simple Bayesian model, a model that used the traditional RW model learning rule, and a model that used the modified RW learning rule, it is possible that other models can account for the findings reported here. The reason we focused on these three models, rather than to focus on other associative and non-associative models that have been used to explain causal reasoning such as the Power PC model (Cheng, 1997; see also Dickinson & Burke, 1996; Jenkins & Ward, 1965), is because these models have been ruled out to be plausible accounts of causal reasoning (for an extensive discussion on this issue, see Griffiths et al., 1995) and because the debate about whether infants, children, and adults can engage in BB reasoning tends to focus on these three models, which make unique predictions about how participants should rate causally redundant cues across the different rating phases. Still, it may be worthwhile in future studies to test the predictions of multiple causal models, including the associative and Bayesian models discussed here, to clarify the nature of the mechanism that underpins causal reasoning.

**Conclusion**

These potential criticisms notwithstanding, these experiments constitute one of the first systematic attempts to examine BB in the context of multiple objects, different contexts, and a series of rating phases. It has been previously proposed that humans use Bayes' rule to reason about causal events that the putative BB finding derives from 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 both an associative mechanism and a Bayesian mechanism may underpin causal reasoning and that whether one mechanism is used over another depends on multiple factors such as the number of potential causes, the context in which learners are asked to make causal inferences, and whether adults are asked to make multiple instance of the same event across multiple rating phases. A fruitful goal of future research will be to examine more closely the claim that whether Bayesian inference is used or associative processing depends on where along "causal gradient' are situated. It may also be worthwhile in future research implement a pre-, mid-, and post-rating phases to assess to what extent children engage in BB reasoning and to use a wider range of potential causal cues than that used here.

Figure 1

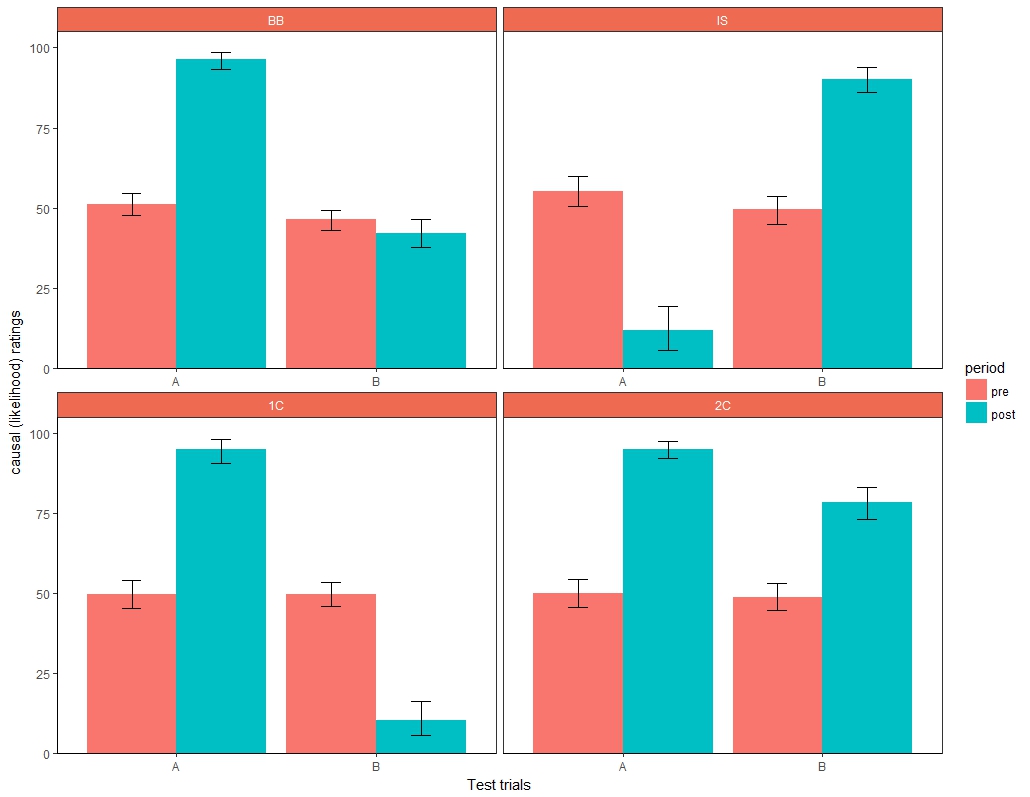


Figure 2

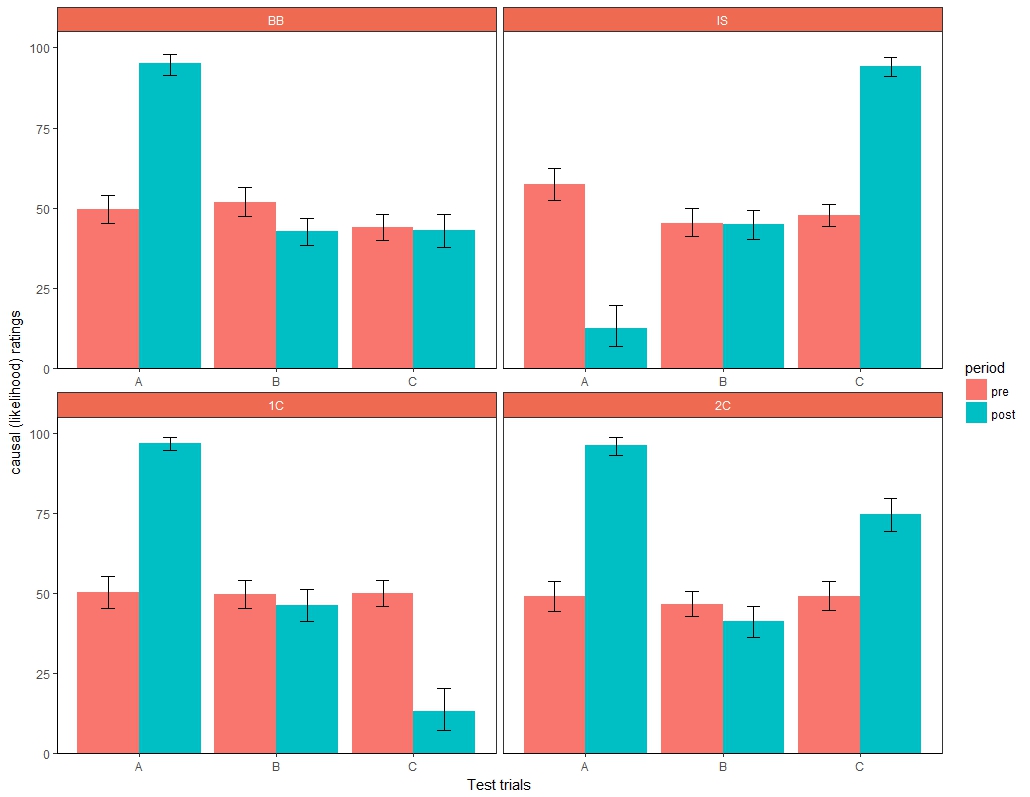


Figure 3

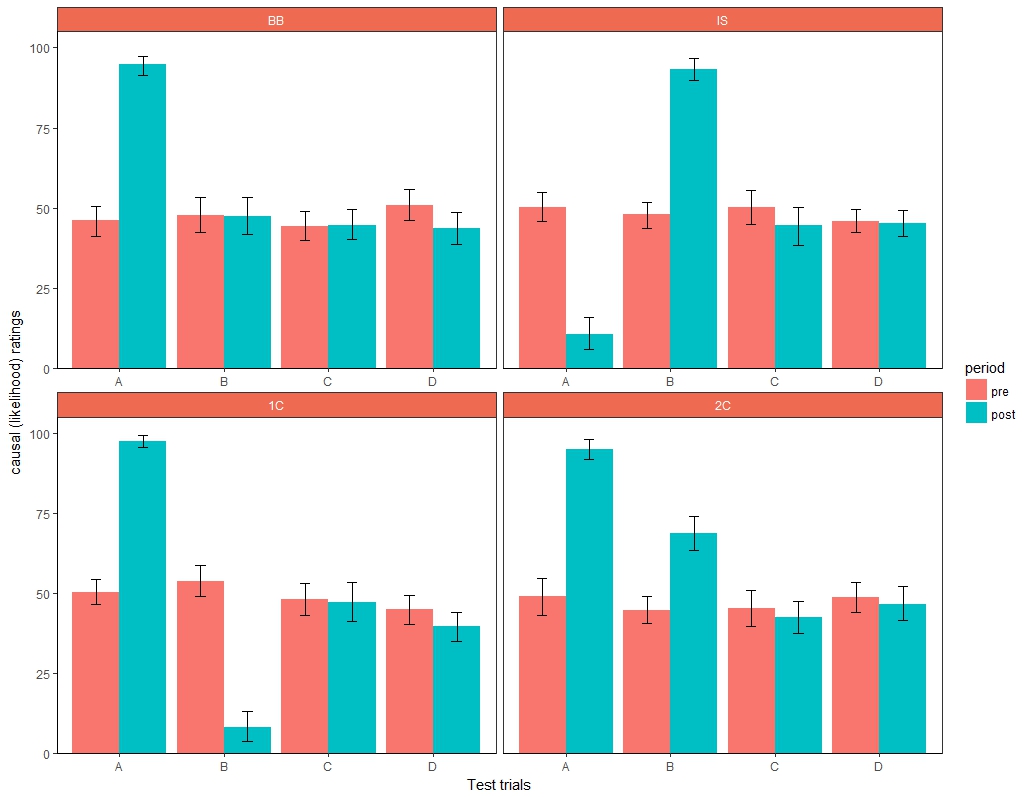


Figure 4

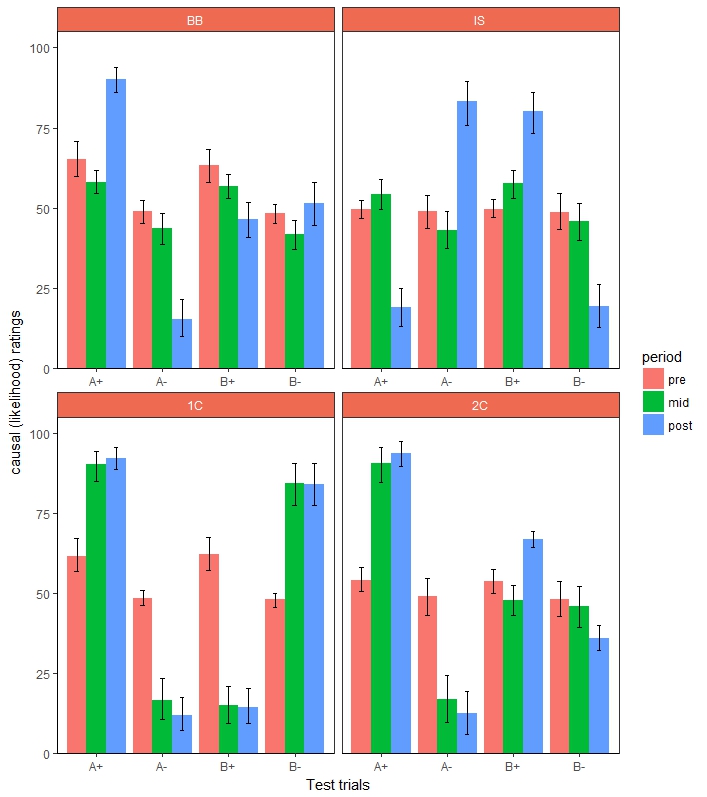


Figure 4



|  |  |  |  |
| --- | --- | --- | --- |
| **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* |

|  |  |  |  |
| --- | --- | --- | --- |
| **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* |

1. Note that we included random effects for participants because models that included random effects yielded smaller Schwartz Bayesian Information Criterion (BIC) values than models that did not include them. [↑](#footnote-ref-1)
2. We used this dummy coding convention throughout the analyses sections in each of the four experiments. [↑](#footnote-ref-2)
3. In each of the analyses presented in Experiments 1-3, one of the β values corresponds to the mean of the reference group (i.e., the intercept in the LMM model) and one of the β values—that is, Δβ—corresponds to the mean difference between the reference and comparison groups. [↑](#footnote-ref-3)