Supplemental Materials for: Extending a Rational Process Model of Causal Reasoning: Assessing Markov Violations and Explaining Away with Inhibitory Causal Relations

Rehder, B.

March 9, 2024

1 Model Fits with Separate Parameters for Exogenous Causes

In the main article, the normative model and mutation sampler were each fit to the results of Experiments 1–3 assuming a single parameter representing the strength of the exogenous causes. w_E represented the exogenous causes of E_A and E_B in the common cause network of Experiment 1, $w_{(Y,Z)}$ represented the exogenous causes of Y and Z in the chain network of Experiment 2, and w_C represented the exogenous causes of C_A and C_B in the common effect network of Experiment 3. Table 1 compares those model fits with those of alternative models in which those exogenous causes were represented by two separate parameters. In all experiments the simpler models achieved a better fit to a larger number of subjects; in only Experiment 1 did the more complex models achieve a better average AIC. The fits presented in the main text for Experiments 1–3 are those for the simpler models.

Table 1: Comparison of models with either one or two parameters representing the strength of exogenous causes.

Experiment	Model Type	Model	Average AIC	% Subjects
Experiment 1	Normative	5-Parameter	124.5	71%
		6-Parameter	124.2	
	Mutation Sampler	6-Parameter	115.9	64%
		7-Parameter	115.4	
Experiment 2	Normative	5-Parameter	135.9	79%
		6-Parameter	136.1	
	Mutation Sampler	6-Parameter	129.4	75%
		7-Parameter	129.9	
Experiment 3	Normative	5-Parameter	138.4	80%
		6-Parameter	139.1	
	Mutation Sampler	6-Parameter	130.3	78%
		7-Parameter	130.3	

2 Fits to All 24 Test Items

In the main article the presentation of the results was limited to those test items that are relevant to Markov violations and (in Experiment 3) explaining away. Figs. 1, 2, and 3 present the fits to all 24 test items in each condition of Experiments 1, 2, and 3, respectively.

3 Sensitivity of Model Fits to Parameter Values

One question that can be asked of the model fits is whether the mutation sampler's advantage holds across a wide range of parameter values or is highly specific to those that yielded the optimal fits. To answer this question for the causal strength and chain length (λ) parameters, the data from each

experiment was fit at the aggregate group (rather than subject) level varying the absolute magnitude of the causal strength parameters from 0.1 to 1.1 in steps of 0.2 and, for the mutation sampler, the chain length parameter across the values 3, 6, 9, 12, 18, 24, 36, 48, and 64. The strength of a causal relation was negated in those conditions in which it was described as inhibitory. (E.g., when fitting a causal strength parameter of 0.5 in Experiment 1's generative-inhibitory condition, w_{CE_A}, the strength of the generative $C \rightarrow E_A$ relationship, was set to 0.5 and w_{CE_B}, the strength of the inhibitory $C \rightarrow E_B$ relationship, was set to -0.5.) The parameters representing the base rate of the causes (e.g., w_C in Experiment 1), the exogenous causes of the effects (e.g., w_E in Experiment 1), and the scaling parameter s remained as free parameters.

Figs. 4, 5, and 6 presents the difference between the AIC values obtained by the mutation sampler and the normative model for each facet in the parameter grid for each condition in Experiments 1, 2, and 3, respectively. Lower AIC values reflect a better fit so negative values in the figures reflect an advantage for the mutation sampler. The figures thus confirm that the mutation sampler yields a better fits in a large portion of the parameter space, not just those parameter values optimized to yield the best fit.

The only corner of the parameter space in which the normative model yields a better fit is when the causal strength magnitudes are high and the chain length parameter is low. Larger causal strengths of course tend to predict more deterministic responding; for the mutation sampler, small chain lengths result in even stronger deterministic responding (because the prototypes reflect deterministic relations, i.e., a cause is always accompanied by its effect). As a result, in this part of the parameter space the advantages that the mutations sampler usually enjoys (e.g., accounting for Markov violations) is offset by its prediction of highly deterministic responding. In contrast, subjects' inferences were much more probabilistic, as reflected by best fitting causal strength parameters with magnitudes typically in the range [0.4, 1.0].

4 Fits of the Interactive Cause Model to Experiment 3

The fits of the interactive cause model described in Appendix B to the results of Experiment 3 are presented in Fig. 7.

5 Fits of the Noisy Logical Models to the results of Experiments 1–3

Figs. 8, 9, and 10 present the fits of the normative model and the mutation sampler assuming noisy logical rather than logistic generating functions to the results of Experiments 1, 2, and 3, respectively.

6 Fits of the Noisy-Logical Model with Separate Parameters for Exogenous Causes

As mentioned in the main text, the fits of the models with noisy-logical generating functions resulted in poor fits (see Fig. 13 in the main text). One approach to address the failure in that figure is to introduce separate parameters, w_{E_A} and w_{E_B}, for the exogenous causes of E_A and E_B, respectively. The resulting fits shown in Fig. 11 confirm that this new model can account for the fact that $p(E_A = 1|C = 0) \ll p(E_B = 1|C = 0)$ in the generative-inhibitory condition and $p(E_A = 1|C = 0) \gg p(E_B = 1|C = 0)$ in the inhibitory-generative condition. However, this model leaves unanswered *why* w_{E_A} and w_{E_B} might differ across conditions. As noted in Appendix A, although the exogenous causes of E_A and E_B might differ when reasoning with real-world materials, the current experimental materials were novel and thus subjects would have no reason to assume they would differ. Moreover, there is no obvious reason why changing the signs of $C \rightarrow E_A$ and $C \rightarrow E_B$ should result in a change in subjects' beliefs about the exogenous causes of E_A and E_B. Yet, the fits in Fig. 11 were achieved only by assuming that subjects believed that the average strength of E_A's exogenous causes was .235 in the generative-inhibitory condition but was .612 in the inhibitory-generative condition. Analogously, the strength of the exogenous causes of E_B was .579 in the generative-inhibitory condition but .255 in the inhibitory-generative condition.

But regardless of whether it is deemed theoretically appropriate, Table 2 compares the fits of the original logistic model (both the 5-parameter normative and 6-parameter mutation sampler versions) against noisy-logical versions with an extra parameter to separately represent the exogenous causes of E_A and E_B in Experiment 1, Y and Z in Experiment 2, and C_A and C_B in Experiment 3. The table reveals that in all experiments the logistic models achieved a better fit to a larger number of subjects; in only Experiment 1 did the noisy-logical models achieve a better average AIC.

Table 2: Fits of the logistic and noisy-logical models to the conditional probability ratings in Experiments 1–3 where the noisy-logical models are allowed an additional parameter to separately represent the strength of the exogenous causes of E_A and E_B in Experiment 1, Y and Z in Experiment 2, and C_A and C_B in Experiment 3.

Experiment	Model Type	Model	Average AIC	% Subjects
Experiment 1	Normative	5-Parameter Logistic	124.5	71%
		6-Parameter Noisy-Logical	124.2	
	Mutation Sampler	6-Parameter Logistic	115.9	58%
		7-Parameter Noisy-Logical	114.7	
Experiment 2	Normative	5-Parameter Logistic	135.9	79%
		6-Parameter Noisy-Logical	140.6	
	Mutation Sampler	6-Parameter Logistic	129.4	82%
		7-Parameter Noisy-Logical	130.2	
Experiment 3	Normative	5-Parameter Logistic	138.4	80%
		6-Parameter Noisy-Logical	141.8	
	Mutation Sampler	6-Parameter Logistic	130.4	72%
		7-Parameter Noisy-Logical	133.7	

7 Predictions of Mental Model Theory

This section derives the predictions of Mental Models Theory (MMT) for the key causal inferences studied in Experiments 1–3. MMT posits that the meanings of causal statements consist of mental representations of concrete states of affairs that together express the possibilities that are allowed by the causal statement. The mental models associated with *cause* and *prevent* are shown in Table 3. Following standard presentations of MMT, this section uses standard Boolean notation throughout, that is, the presence and absence of variable X, denoted $X = 1$ and $X = 0$ in the rest of the paper, are denoted with lower case x and $\neg x$ here.

According to Johnson-Laird and Byrne (2002), the same causal verb can be associated with different sets of possibilities. Because of working memory limitations, a reasoner will often represent a causal scenario with an *initial* set of mental models and draw causal inferences on their basis. The initial models are those in which the variables states explicitly mentioned in the casual statements (c and e in Table 3) are assumed to be true. The ellipses represent additional *implicit* possibilities—ones the reasoner knows exist but does not represent explicitly. Examples of how ellipses influence the computation of conditional probabilities (which is also moderated by which variables in the explicit initial models that are bracketed, e.g., [c] in Table 3) are provided below.

In contrast, *fully explicit* mental models are those that represent the complete set of possibilities. There are weak and strong interpretations of both *cause* and *prevent*. The weak interpretation of *cause* implies sufficiency but not necessity (i.e., the effect has other potential causes) whereas the strong interpretation also implies necessity (and thus rules out $\neg ce$ as a possibility). Analogously, the weak interpretation of *prevent* implies other potential preventers whereas the strong version does not (ruling out $\neg c \neg e$ as a possibility). Which interpretation is chosen depends on the background knowledge that reasoners bring to a particular causal scenario. However, note that in all interpretations MMT stipulates causal relations that are deterministically sufficient: a cause always produces an effect (and thus $c\neg e$ never appears as a possibility) and a preventer always prevents it (and thus CE never appears as a possibility).

The representation of larger causal networks are constructed from these primitive representations. For example, for the common cause network $E_1 \leftarrow C \rightarrow E_2$, one takes the Cartesian product of the mental models for $C \rightarrow E_1$ and for $C \rightarrow E_2$ and then eliminates contradictions and duplicates.

Table 3: Mental models for *cause* and *prevent*.

	C *causes* E *("generative")*		C *prevents* E *("inhibitory")*	
Initial	[c]	e	[c]	$\neg e$
	...		...	
Fully Explicit	c	e	c	$\neg e$
(Weak Interpretation)	$\neg c$	e	$\neg c$	e
	$\neg c$	$\neg e$	$\neg c$	$\neg e$
Fully Explicit	c	e	c	$\neg e$
(Strong Interpretation)	$\neg c$	$\neg e$	$\neg c$	e

Assuming the fully explicit weak interpretation of *cause* shown in Table 3, the mental models for $E_1 \leftarrow C \rightarrow E_2$ are $\{ce_1, \neg ce_1, \neg c\neg e_1\} \times \{ce_2, \neg ce_2, \neg c\neg e_2\}$, which, after contradictions are removed, yields $\{ce_1e_2, \neg ce_1e_2, \neg c\neg e_1e_2, \neg ce_1\neg e_2, \neg ce_1\neg e_2\}$.

Conditional probabilities are computed by first selecting those models in which the antecedent is true and then calculating the proportion of those models in which the consequent is true. For the weak common cause example above, to compute $p(e_1|e_2)$ one identifies all models that include e_2, namely, $\{ce_1e_2, \neg ce_1e_2, \neg c\neg e_1e_2\}$. e_1 appears in two of those three models; thus $p(e_1|e_2) = .67$.

The sections below present MMT's predictions for common cause, chain, and common effect networks, both when both causal relations are generative and when one is generative and the other inhibitory. These analyses follow ones previously presented in Ali et al. (2011), which in turn were based on presentations of MMT in Byrne et al. (1999), Goldvarg and Johnson-Laird (2001), Johnson-Laird and Byrne (2002), and Johnson-Laird et al. (1999). Here Ali et al.'s analyses have been extended to preventative in addition to generative relations, the strong interpretations of the causal relations in addition to the weak ones, and to chain networks in addition to common cause and common effect networks.

7.1 Predictions for a Generative-Generative Common Cause Network

Table 4 presents the MMT predictions for a common cause network in which both causal relations are generative (what this article has labeled the generative-generative condition). The three sections of the table evaluate MMT's predictions for the initial mental models, the fully explicit models under the weak interpretation, and the fully explicit models under the strong interpretation. The first column presents the mental models for a common cause network under each interpretation, formed by combining the appropriate interpretations of $C \rightarrow E_1$ and $C \rightarrow E_2$ shown in Table 3. The second column shows the mental models that result by conditioning on c, e_2, and ce_2. The remaining columns show the conditional probabilities that result.

Columns 3-5 present MMT's predictions regarding the unconditional dependence of E_1 and E_2 by comparing its predictions for $p(e_1)$ and $p(e_1|e_2)$[1]. (As throughout this article, the "unconditional dependence" between E_1 and E_2 refers to the relationship between E_1 and E_2 given no knowledge of the state of C). According to the normative causal graphical model, in the absence of any information about C, E_1 and E_2 are dependent such that, when both causal relations are generative, the presence of e_2 should make e_1 more likely. Recall that in MMT conditional probabilities are computed by calculating, among the models in which the antecedent is true, the proportion of models in which the consequent is also true. Starting with fully explicit weak models, $p(e_1) = .60$ because, of the five common cause models shown in the first column, e_1 is true in three of them. $p(e_1|e_2) = .67$, because in the three models that result from conditioning on e_2, e_1 is true in two of them. Because $p(e_1|e_2) > p(e_1)$, MMT correctly predicts the positive dependency between e_1 and e_2 in a common cause network with generative relations. Table 4 shows that this qualitative prediction also holds for the strong interpretation of the causal relations.

Computing conditional probabilities for the initial models requires special handling for the ellipses.

[1]

Table 4: Mental model theory's predictions for key inferences in a common cause network in which $C \rightarrow E_1$ and $C \rightarrow E_2$ are both generative.

Initial Models

Mental Models			Models After Conditioning			$p(e_1)$	$p(e_1\|e_2)$	Unconditional Dependence	$p(e_1\|c)$	$p(e_1\|ce_2)$	Screening Off
[c]	e_1	e_2	Given c:			0.50	0.50	0	1.0	1.0	0*
...			c	e_1	e_2						
			Given e_2:								
			[c]	e_1	e_2						
			...								
			Given ce_2:								
			c	e_1	e_2						

Fully Explicit Models (Weak Interpretation)

Mental Models			Models After Conditioning			$p(e_1)$	$p(e_1\|e_2)$	Unconditional Dependence	$p(e_1\|c)$	$p(e_1\|ce_2)$	Screening Off
c	e_1	e_2	Given c:			0.60	0.67	> 0*†	1.0	1.0	0*
$\neg c$	e_1	e_2	c	e_1	e_2						
$\neg c$	e_1	$\neg e_2$	Given e_2:								
$\neg c$	$\neg e_1$	e_2	c	e_1	e_2						
$\neg c$	$\neg e_1$	$\neg e_2$	$\neg c$	e_1	e_2						
			$\neg c$	$\neg e_1$	e_2						
			Given ce_2:								
			c	e_1	e_2						

Fully Explicit Models (Strong Interpretation)

Mental Models			Models After Conditioning			$p(e_1)$	$p(e_1\|e_2)$	Unconditional Dependence	$p(e_1\|c)$	$p(e_1\|ce_2)$	Screening Off
c	e_1	e_2	Given c:			0.50	1.0	> 0*†	1.0	1.0	0*
$\neg c$	$\neg e_1$	$\neg e_2$	c	e_1	e_2						
			Given e_2:								
			c	e_1	e_2						
			Given ce_2:								
			c	e_1	e_2						

Note. Unconditional dependence = $p(e_1|e_2) - p(e_1)$. Screening off = $p(e_1|ce_2) - p(e_1|c)$. * Prediction is qualitatively consistent with the normative model. † Prediction is qualitatively consistent with the empirical results observed in most studies.

Initial models represent those possibilities in which the variables explicitly mentioned as part of the premises (in the common cause case, c, e_1, and e_2) are assumed to be true. The ellipses are the reasoner's summary (i.e., implicit) representation of possibilities in which some of those variables are false. Thus, in the computation of conditional probabilities the ellipses are generally treated as a (single) instance in which the variable whose probability is being queried is false. (See below for an exception to this handling of ellipsis.) Thus, $p(e_1) = .50$, because of the two common cause models show in the first column (including the ellipses), e_1 is true in one of them. Because the probability e_1 when conditioning on e_2 is also .50, the difference between $p(e_1)$ and $p(e_1|e_2)$ is 0, and so the initial models do not predict the positive dependency between e_1 and e_2 required by a common cause network with generative relations.

Columns 7-9 of Table 4 show MMT's predictions regarding the *conditional* dependence of E_1 and E_2 (i.e., conditioned on C) in a common cause network with generative relations by comparing MMT's predictions for $p(e_1|c)$ and $p(e_1|ce_2)$. According to the normative model, in the presence of information about C, E_1 and E_2 should be independent, that is, E_1 and E_2 are screened off from each other by C. Table 4 shows that MMT predicts screening off under all interpretations of the causal relations, consistent with the normative model. Of course, this prediction also means that MMT fails to predict the violations of independence exhibited by human reasoners when reasoning with a common cause network that have been widely reported in the literature and the current Experiment 1.

Table 5: Mental model theory's predictions for key inferences in a common cause network in which $C \rightarrow E_1$ is inhibitory and $C \rightarrow E_2$ is generative.

Initial Models

Mental Models	Models After Conditioning	$p(e_1)$	$p(e_1\|e_2)$	Unconditional Dependence	$p(e_1\|c)$	$p(e_1\|ce_2)$	Screening Off
$[c]$ $\neg e_1$ e_2	Given c:	0.50	0.50	0	0	0	0*
...	c $\neg e_1$ e_2						
	Given e_2:						
	$[c]$ $\neg e_1$ e_2						
	...						
	Given ce_2:						
	c $\neg e_1$ e_2						

Fully Explicit Models (Weak Interpretation)

Mental Models	Models After Conditioning	$p(e_1)$	$p(e_1\|e_2)$	Unconditional Dependence	$p(e_1\|c)$	$p(e_1\|ce_2)$	Screening Off
c $\neg e_1$ e_2	Given c:	0.40	0.33	< 0*†	0	0	0*
$\neg c$ e_1 e_2	c $\neg e_1$ e_2						
$\neg c$ e_1 $\neg e_2$	Given e_2:						
$\neg c$ $\neg e_1$ e_2	c $\neg e_1$ e_2						
$\neg c$ $\neg e_1$ $\neg e_2$	$\neg c$ e_1 e_2						
	$\neg c$ $\neg e_1$ e_2						
	Given ce_2:						
	c $\neg e_1$ e_2						

Fully Explicit Models (Strong Interpretation)

Mental Models	Models After Conditioning	$p(e_1)$	$p(e_1\|e_2)$	Unconditional Dependence	$p(e_1\|c)$	$p(e_1\|ce_2)$	Screening Off
c $\neg e_1$ e_2	Given c:	0.50	0	< 0*†	0	0	0*
$\neg c$ e_1 $\neg e_2$	c e_1 e_2						
	Given e_2:						
	c $\neg e_1$ e_2						
	Given ce_2:						
	c $\neg e_1$ e_2						

Note. Unconditional dependence = $p(e_1|e_2) - p(e_1)$. Screening off = $p(e_1|ce_2) - p(e_1|c)$. * Prediction is qualitatively consistent with the normative model. † Prediction is qualitatively consistent with the empirical results observed in Experiment 1.

7.2 Predictions for an Inhibitory-Generative Common Cause Network

Table 5 presents MMT's predictions for a common cause network in which $C \rightarrow E_1$ is inhibitory and $C \rightarrow E_2$ is generative, labeled as the inhibitory-generative condition in the main article. The results are very analogous to those in the generative-generative condition described in Appendix D. With the exception of the initial models, MMT correctly predicts unconditional dependence between E_1 and E_2, albeit E_2 now makes E_1 *less* likely. This prediction is consistent with both the normative model and human reasoners in Experiment 1's inhibitory-generative condition. However, as it did in the generative-generative condition MMT again predicts screening off, which is consistent with the normative model but not the judgments of Experiment 1's inhibitory-generative subjects, who violated independence.

In summary, for a common cause network MMT predicts (although not for initial models) that E_1 and E_2 will exhibit unconditional dependence in a direction that is consistent with the normative model and human reasoners. However, MMT is unable to account for the pervasive presence of screening off errors that humans exhibit when judging the contingency between E_1 and E_2 conditioned on C.

7.3 Predictions for a Generative-Generative Chain Network

Table 6 presents MMT's predictions for the chain network $X \rightarrow Y \rightarrow Z$ with generative relations. Table 6 present predictions for the initial mental models and the weak and strong interpretations of the fully explicit models. The first column presents the models for a chain network under each interpretation, formed by combining the appropriate interpretations of $X \rightarrow Y$ and $Y \rightarrow Z$ from Table 12 in the main article. The second column shows the mental models that result by conditioning on y, x, and xy.

Table 6 shows that MMT correctly predicts the positive unconditional dependency between X and Z—specifically, that $p(z|x) > p(z)$—regardless of whether one assumes the initial models or the two fully explicit representations. Regarding the conditional dependency between X and Z conditioned on Y, MMT's fully explicit mental models predict that Y screens off Z from X. However, as was the case with the common cause network, although this prediction agrees with the normative model, it disagrees with the behavior of human reasoners who violate screening off in chain networks (e.g., in the current Experiment 2).

Interestingly, the initial models do predict that $p(z|xy) > p(z|y)$, that is, a violation of screening off in agreement with subject's behavior. However, the conditional dependency between X and Z can be assessed not only by examining the conditional probability of Z given various values of X and Y but also by the conditional probability of X given values of Y and Z. To this end, the initial models section of Table 6 also presents MMT's predictions for $p(x)$, $p(x|\neg z)$, $p(x|y)$ and $p(x|y\neg z)$. For *these* inferences, MMT fails to predict the positive unconditional dependency between X and Z, namely, that $p(x|\neg z)$ should be $> p(x)$. And, it predicts screening away: that $p(x|y\neg z) = p(x|y)$. In contrast, Fig. 2A reveals that subjects in Experiment 2's generative-generative condition violate independence when predicting x given y and z just as they do when predicting z given x and y.

7.4 Predictions for an Inhibitory-Generative Chain Network

Table 7 present MMT's predictions for the chain network $X \rightarrow Y \rightarrow Z$ in which $X \rightarrow Y$ is inhibitory and $Y \rightarrow Z$ is generative. The results are analogous to those in the generative-generative condition described above: Most representations predict the (now negative) unconditional dependency between X and Z (consistent with the normative model and human subjects) and the absence of screening off violations (consistent with the normative model but not human subjects). Once again, the screening off results for initial models depend on whether one is predicting the conditional probability of X given values for Y and Z or Z given values for X and Y (i.e., $p(z|xy) < p(z|y)$ but $p(x|y\neg z) = p(x|y)$).

In summary, as was the case for the common cause networks, MMT does not provide a comprehensive account of the violations of screening off that occur with chain networks. MMT correctly predicts that X and Z will exhibit unconditional dependence in a direction that is consistent with the normative model and human reasoners, although not for initial models when dependence is assessed by computing the conditional probability of X given values for Y and Z.

7.5 Predictions for a Generative-Generative Common Effect Network

Table 8 presents MMT's predictions for the common effect network $C_1 \rightarrow E \leftarrow C_2$ with generative.

The first column presents the initial and fully explicit (weak and strong interpretations) formed by combining the appropriate interpretations of $C_1 \rightarrow E$ and $C_2 \rightarrow E$ from Table 3.

The second column shows the mental models that result by conditioning on e, c_2, and c_2E. Note that the initial models are those that result when background knowledge informs the reasoner that C_1 and C_2 have independent causal influences on E (Goldvarg et al., 1999).

In the current Experiment 3, independence was communicated to subjects by presenting them with separate descriptions of the causal mechanisms associated with $C_1 \rightarrow E$ and $C_2 \rightarrow E$ that implied that they were independent.

The table shows that, for a common effect network with generative relations, MMT's fully explicit models yield a positive unconditional dependency between C_1 and C_2. This differs from the normative model, which stipulates that C_1 and C_2 are unconditionally independent, but conforms with the judgments of human reasoners who violate independence (e.g., in Experiment 3's generative-generative condition and numerous other studies). Regarding the conditional dependency between C_1 and C_2, Table 8 reveals that none of the models yield explaining away. Although the current Experiment 3's

Table 6: Mental model theory's predictions for key inferences in a chain network in which $X \rightarrow Y$ and $Y \rightarrow Z$ are both generative.

Initial Models

Mental Models			Models After Conditioning	$p(z)$	$p(z\|x)$	Unconditional Dependence	$p(z\|y)$	$p(z\|xy)$	Screening Off
$[x]$	y	z	Given y:	0.50	1.0	> 0*†	0.50	1.0	> 0†
...			$[x]$ y z						
			...						
			Given x:						
			x y z						
			Given xy:						
			x y z			Unconditional			Screening
			Given: z	$p(x)$	$p(x\|z)$	Dependence	$p(x\|y)$	$p(x\|yz)$	Off
			$[x]$ y z	0.50	0.50	0	0.50	0.50	0*
			...						
			Given: YZ						
			$[x]$ y z						
			...						

Fully Explicit Models (Weak Interpretation)

Mental Models			Models After Conditioning	$p(z)$	$p(z\|x)$	Unconditional Dependence	$p(z\|y)$	$p(z\|xy)$	Screening Off
x	y	z	Given y:	0.75	1.0	> 0*†	1.0	1.0	0*
$\neg x$	y	z	x y z						
$\neg x$	$\neg y$	z	$\neg x$ y z						
$\neg x$	$\neg y$	$\neg z$	Given x:						
			x y z						
			Given xy:						
			x y z						

Fully Explicit Models (Strong Interpretation)

Mental Models			Models After Conditioning	$p(z)$	$p(z\|x)$	Unconditional Dependence	$p(z\|y)$	$p(z\|xy)$	Screening Off
x	y	z	Given y:	0.50	1.0	> 0*†	1.0	1.0	0*
$\neg x$	$\neg y$	$\neg z$	x y z						
			Given x:						
			x y z						
			Given xy:						
			x y z						

Note. Unconditional dependence = $p(z|x) - p(z)$ and $p(x|z) - p(x)$. Screening off = $p(z|xy) - p(z|y)$ and $p(x|YZ) - p(x|y)$. * Prediction is qualitatively consistent with the normative model. † Prediction is qualitatively consistent with the empirical results observed in most studies.

Table 7: Mental model theory's predictions for key inferences in a chain network in which $X \rightarrow Y$ is generative and $Y \rightarrow Z$ is inhibitory.

Initial Models

Mental Models	Models After Conditioning	$p(z)$	$p(z\|x)$	Unconditional Dependence	$p(z\|y)$	$p(z\|xy)$	Screening Off
$[x]$ y $\neg z$	Given y:	0.50	0	< 0*†	0.50	0	< 0†
...	$[x]$ y $\neg z$						
	...						
	Given x:						
	x y $\neg z$						
	Given xy:						
	x y $\neg z$						
	Given: $\neg z$	$p(x)$	$p(x\|\neg z)$	Unconditional Dependence	$p(x\|y)$	$p(x\|y\neg z)$	Screening Off
	$[x]$ y $\neg z$	0.50	0.50	0	0.50	0.50	0*
	...						
	Given: $y\neg z$						
	$[x]$ y $\neg z$						
	...						

Fully Explicit Models (Weak Interpretation)

Mental Models	Models After Conditioning	$p(z)$	$p(z\|x)$	Unconditional Dependence	$p(z\|y)$	$p(z\|xy)$	Screening Off
x y $\neg z$	Given y:	0.25	0	< 0*†	0	0	0*
$\neg x$ y $\neg z$	x y $\neg z$						
$\neg x$ $\neg y$ z	$\neg x$ y $\neg z$						
$\neg x$ $\neg y$ $\neg z$	Given x:						
	x y $\neg z$						
	Given xy:						
	x y $\neg z$						

Fully Explicit Models (Strong Interpretation)

Mental Models	Models After Conditioning	$p(z)$	$p(z\|x)$	Unconditional Dependence	$p(z\|y)$	$p(z\|xy)$	Screening Off
x y $\neg z$	Given y:	0.50	0	< 0*†	0	0	0*
$\neg x$ $\neg y$ z	x y $\neg z$						
	Given x:						
	x y $\neg z$						
	Given xy:						
	x y $\neg z$						

Note. Unconditional dependence = $p(z|x) - p(z)$ and $p(x|\neg z) - p(x)$. Screening off = $p(z|xy) - p(z|y)$ and $p(x|y\neg z) - p(x|y)$. * Prediction is qualitatively consistent with the normative model. † Prediction is qualitatively consistent with the empirical results observed in Experiment 2.

generative-generative condition and other studies reveal that explaining away is often weaker than it should be, human reasoner do explain away, a normative causal inference that MMT is unable to reproduce.

7.6 Predictions for an Inhibitory-Generative Common Effect Network

Table 9 present MMT's predictions for the common effect network $C_1 \rightarrow E \leftarrow C_2$ in which in which $C_1 \rightarrow E$ is inhibitory and $C_2 \rightarrow E$ is generative. The results are analogous to those in the generative-generative common effect condition: Although, the fully explicit representations predict the (now negative) unconditional dependency between C_A and C_B, none of the representations predict explaining away, as exhibited by the normative model and human reasoners.

In summary, whereas the fully explicit representations predict the unconditional dependency between C_A and C_B, none of the representation predict explaining away, as exhibited by the normative model and the human reasoners. Thus, as was the case in for common cause and chain networks, MMT does not provide a complete account of how humans reason with common effect networks.

8 Examples of the Experimental Interface

Figs. 12– 13 present screenshots from the experimental interface, including the verbal presentation of the causal relations (Fig. 12), a diagram of those relations (Fig. 13), an example from the multiple-choice question test that subjects were required to pass (Fig. 14), and an inference question in which subjects were asked for a conditional probability judgment (Fig. 15).

References

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403–418.

Byrne, R., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inference. *Journal of Memory and Language*, *40*, 347–373.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning1. *Cognitive Science*, *25*(4), 565–610.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, J. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*(1), 62–88.

Table 8: Mental model theory's predictions for key inferences in a common effect network in which $C_1 \rightarrow E$ and $C_2 \rightarrow E$ are both generative.

Initial Models

Mental Models	Models After Conditioning	$p(c_1)$	$p(c_1\|c_2)$	Unconditional Dependence	$p(c_1\|e)$	$p(c_1\|c_2e)$	Explaining Away
$[c_1 \quad c_2] \quad e$	Given e:	0.50	0.50	0*	0.50	0.50	0
$[c_1] \quad e$	$[c_1 \quad c_2] \quad e$						
$[c_2] \quad e$	$[c_1] \quad e$						
...	$[c_2] \quad e$						
	...						
	Given c_2:						
	$c_1 \quad c_2 \quad e$						
	$c_2 \quad e$						
	Given c_2e:						
	$c_1 \quad c_2 \quad e$						
	$c_2 \quad e$						

Fully Explicit Models (Weak Interpretation)

Mental Models	Models After Conditioning	$p(c_1)$	$p(c_1\|c_2)$	Unconditional Dependence	$p(c_1\|e)$	$p(c_1\|c_2e)$	Explaining Away
$c_1 \quad c_2 \quad e$	Given e:	0.40	0.50	> 0†	0.50	0.50	0
$c_1 \quad \neg c_2 \quad e$	$c_1 \quad c_2 \quad e$						
$\neg c_1 \quad c_2 \quad e$	$c_1 \quad \neg c_2$						
$\neg c_1 \quad \neg c_2 \quad e$	$\neg c_1 \quad c_2 \quad e$						
$\neg c_1 \quad \neg c_2 \quad \neg e$	$\neg c_1 \quad \neg c_2 \quad e$						
	Given c_2:						
	$c_1 \quad c_2 \quad e$						
	$\neg c_1 \quad c_2 \quad e$						
	Given c_2e:						
	$c_1 \quad c_2 \quad e$						
	$\neg c_1 \quad c_2 \quad e$						

Fully Explicit Models (Strong Interpretation)

Mental Models	Models After Conditioning	$p(c_1)$	$p(c_1\|c_2)$	Unconditional Dependence	$p(c_1\|e)$	$p(c_1\|c_2e)$	Explaining Away
$c_1 \quad c_2 \quad e$	Given e:	0.50	1.0	> 0†	1.0	1.0	0
$\neg c_1 \quad \neg c_2 \quad \neg e$	$c_1 \quad c_2 \quad e$						
	Given c_2:						
	$c_1 \quad c_2 \quad e$						
	Given c_2e:						
	$c_1 \quad c_2 \quad e$						

Note. Unconditional independence $= p(c_1|c_2) - p(c_1)$. Explaining away $= p(c_1|c_2e) - p(c_1|e)$. * Prediction is qualitatively consistent with the normative model. † Prediction is qualitatively consistent with the empirical results observed in most studies.

Table 9: Mental model theory's predictions for key inferences in a common effect network in which $C_1 \rightarrow E$ is inhibitory and $C_2 \rightarrow E$ is generative.

Initial Models

Mental Models			Models After Conditioning	$p(c_1)$	$p(c_1\|c_2)$	Unconditional Dependence	$p(c_1\|e)$	$p(c_1\|c_2e)$	Explaining Away
$[c_1$	$c_2]$	$\neg e$	Given e:	0.50	0.50	0	0	0	0
$[c_1]$		$\neg e$	$[c_2]$ e						
	$[c_2]$	e	...						
...			Given c_2:						
			c_1 c_2 $\neg e$						
			c_2 e						
			Given c_2e:						
			c_2 e						

Fully Explicit Models (Weak Interpretation)

Mental Models			Models After Conditioning	$p(c_1)$	$p(c_1\|c_2)$	Unconditional Dependence	$p(c_1\|e)$	$p(c_1\|c_2e)$	Explaining Away
c_1	$\neg c_2$	$\neg e$	Given e:	0.25	0	< 0†	0	0	0
$\neg c_1$	c_2	e	$\neg c_1$ c_2 e						
$\neg c_1$	$\neg c_2$	e	$\neg c_1$ $\neg c_2$ e						
$\neg c_1$	$\neg c_2$	$\neg e$	Given c_2:						
			$\neg c_1$ c_2 e						
			Given c_2e:						
			$\neg c_1$ c_2 e						

Fully Explicit Models (Strong Interpretation)

Mental Models			Models After Conditioning	$p(c_1)$	$p(c_1\|c_2)$	Unconditional Dependence	$p(c_1\|e)$	$p(c_1\|c_2e)$	Explaining Away
$\neg c_1$	c_2	e	Given e:	0.50	0	< 0†	0	0	0
c_1	$\neg c_2$	$\neg e$	$\neg c_1$ c_2 e						
			Given c_2:						
			$\neg c_1$ c_2 e						
			Given c_2e:						
			$\neg c_1$ c_2 e						

Note. Unconditional independence $= p(c_1|c_2) - p(c_1)$. Explaining away $= p(c_1|c_2e) - p(c_1|e)$. * Prediction is qualitatively consistent with the normative model. † Prediction is qualitatively consistent with the empirical results observed in Experiment 3.

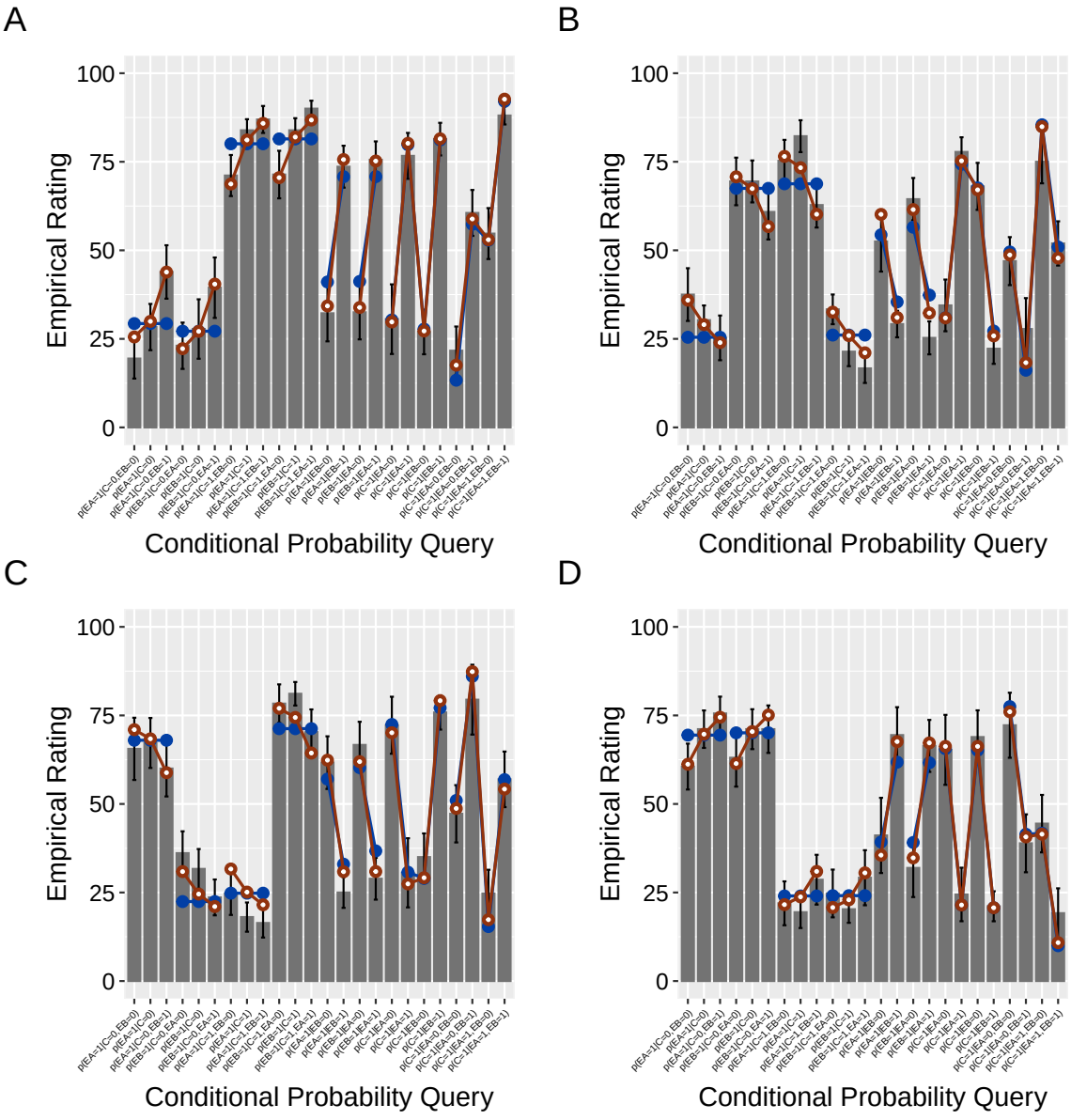

Figure 1: Fits to all test items in Experiment 1. A. Generative-generative condition. B. Generative-inhibitory condition. C. Inhibitory-generative condition. D. Inhibitory-inhibitory condition. The blue and red lines are the fits of the normative model and the mutation sampler, respectively. Error bars are 95% confidence intervals.

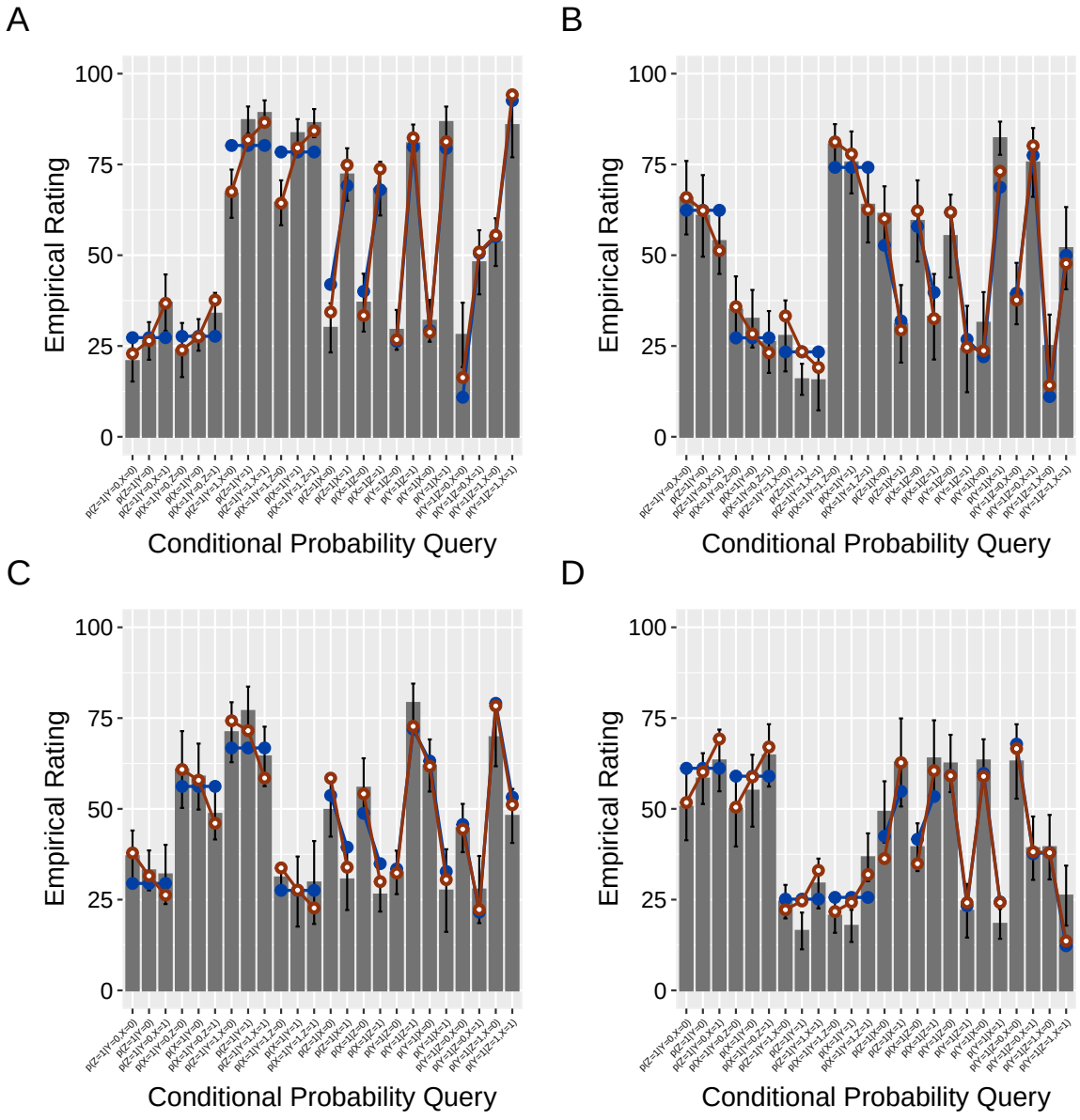

Figure 2: Fits to all test items in Experiment 2. A. Generative-generative condition. B. Generative-inhibitory condition. C. Inhibitory-generative condition. D. Inhibitory-inhibitory condition. The blue and red lines are the fits of the normative model and the mutation sampler, respectively. Error bars are 95% confidence intervals.

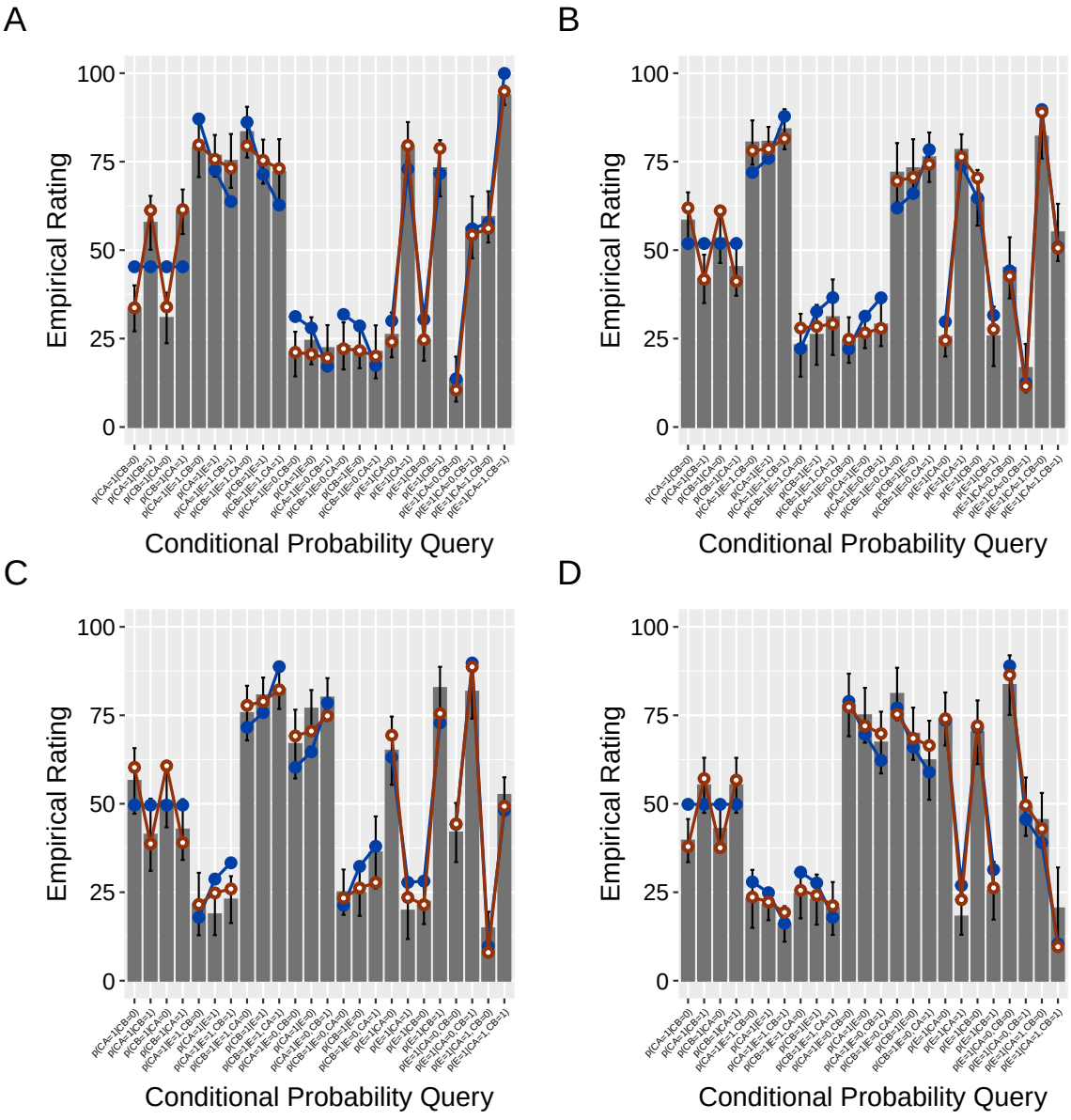

Figure 3: Fits to all test items in Experiment 3. A. Generative-generative condition. B. Generative-inhibitory condition. C. Inhibitory-generative condition. D. Inhibitory-inhibitory condition. The blue and red lines are the fits of the normative model and the mutation sampler, respectively. Error bars are 95% confidence intervals.

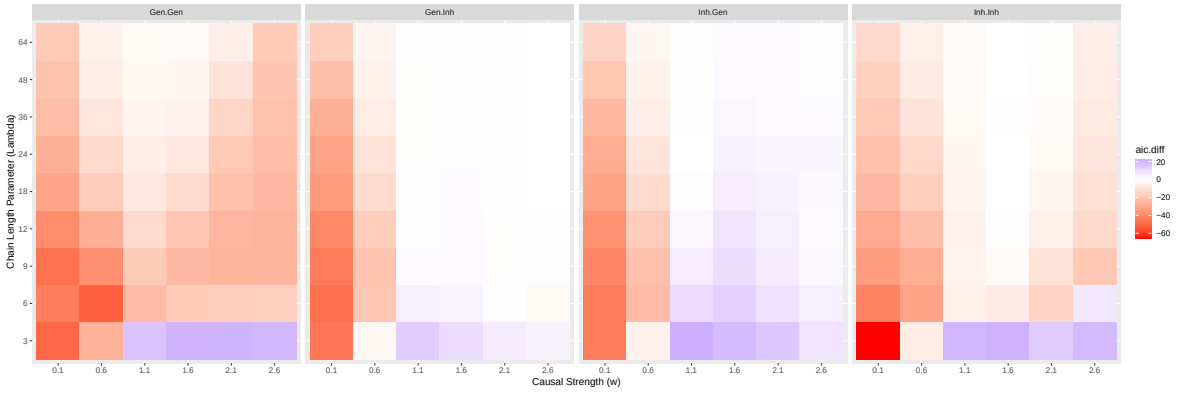

Figure 4: The difference between the fitted AIC values of the mutation sampler and those of the normative model for a range of causal strength and chain length λ parameter values in each condition of Experiment 1. Negative values represent an advantage for the mutation sampler.

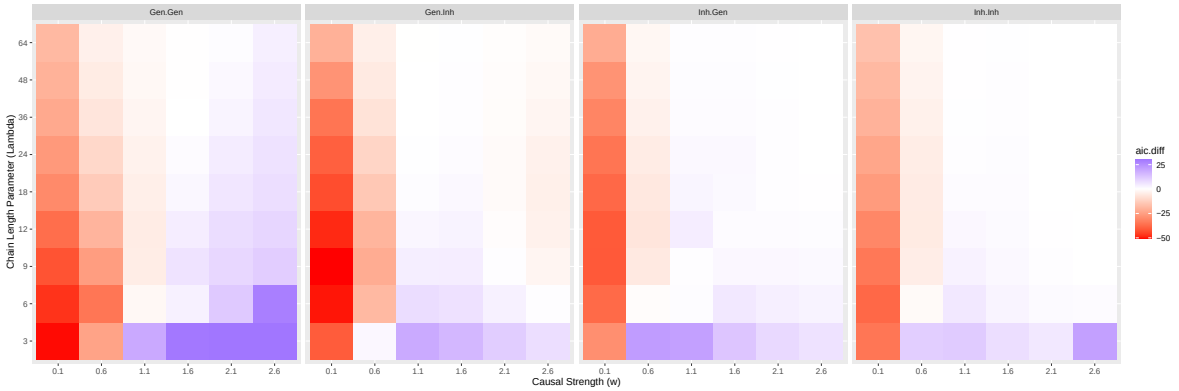

Figure 5: The difference between the fitted AIC values of the mutation sampler and those of the normative model for a range of causal strength and chain length λ parameter values in each condition of Experiment 2. Negative values represent an advantage for the mutation sampler.

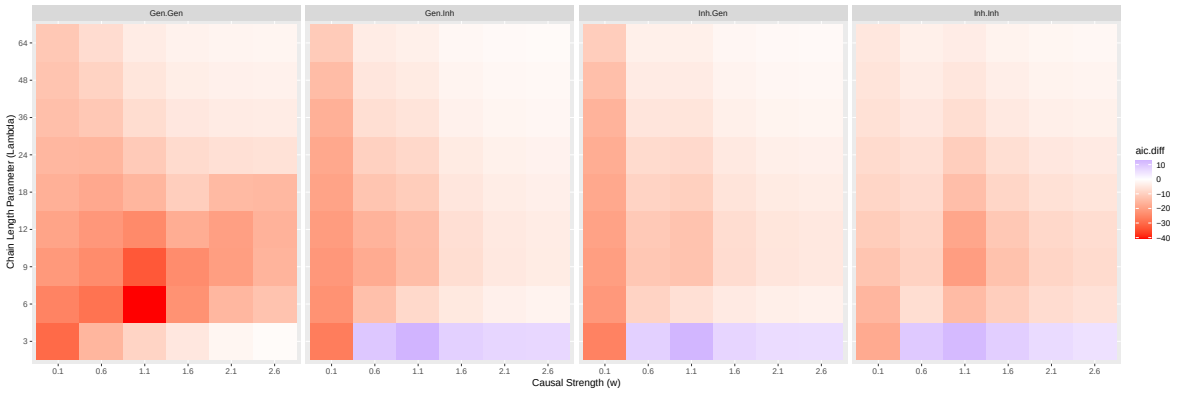

Figure 6: The difference between the fitted AIC values of the mutation sampler and those of the normative model for a range of causal strength and chain length λ parameter values in each condition of Experiment 3. Negative values represent an advantage for the mutation sampler.

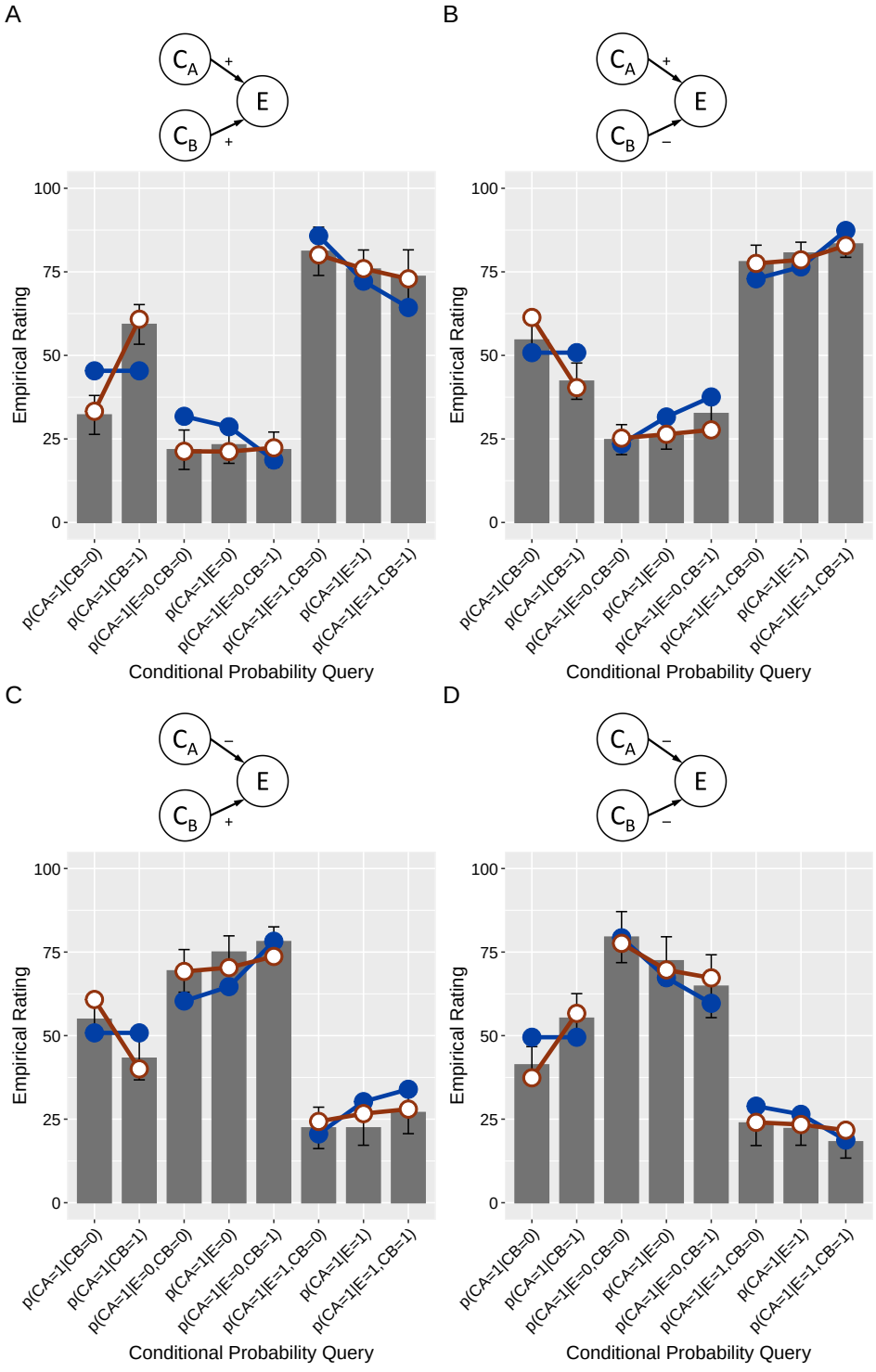

Figure 7: Fits of the interactive cause model (both normative and mutation sampler versions) to the data of Experiment 3. A. Generative-generative condition. B. Generative-inhibitory condition. C. Inhibitory-generative condition. D. Inhibitory-inhibitory condition. The blue and red lines are the fits of the normative model and the mutation sampler, respectively. Error bars are 95% confidence intervals.

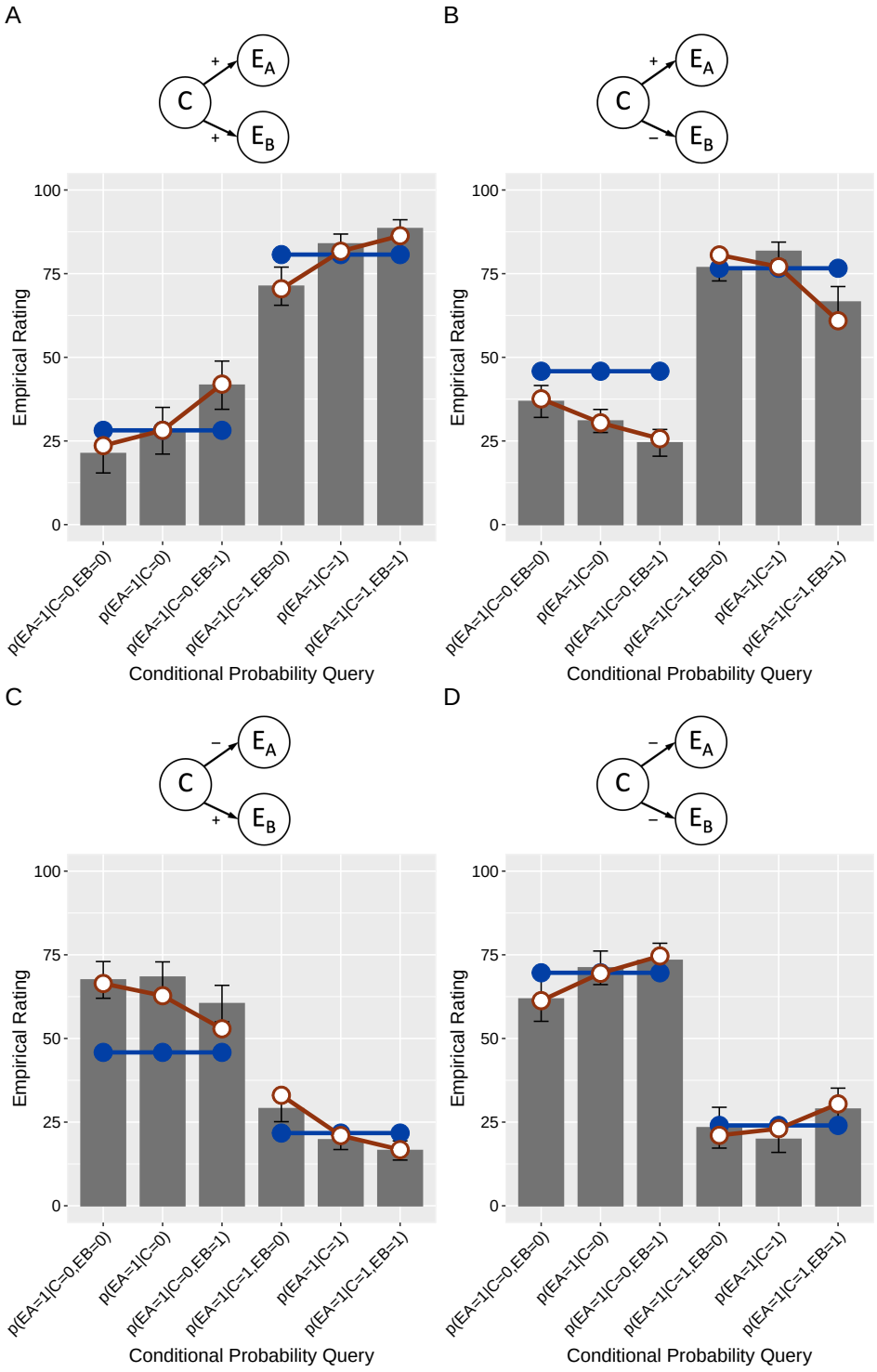

Figure 8: Noisy logical fits to tests items in Experiment 1. A. Generative-generative condition. B. Generative-inhibitory condition. C. Inhibitory-generative condition. D. Inhibitory-inhibitory condition. The blue and red lines are the fits of the normative model and the mutation sampler, respectively. Error bars are 95% confidence intervals.

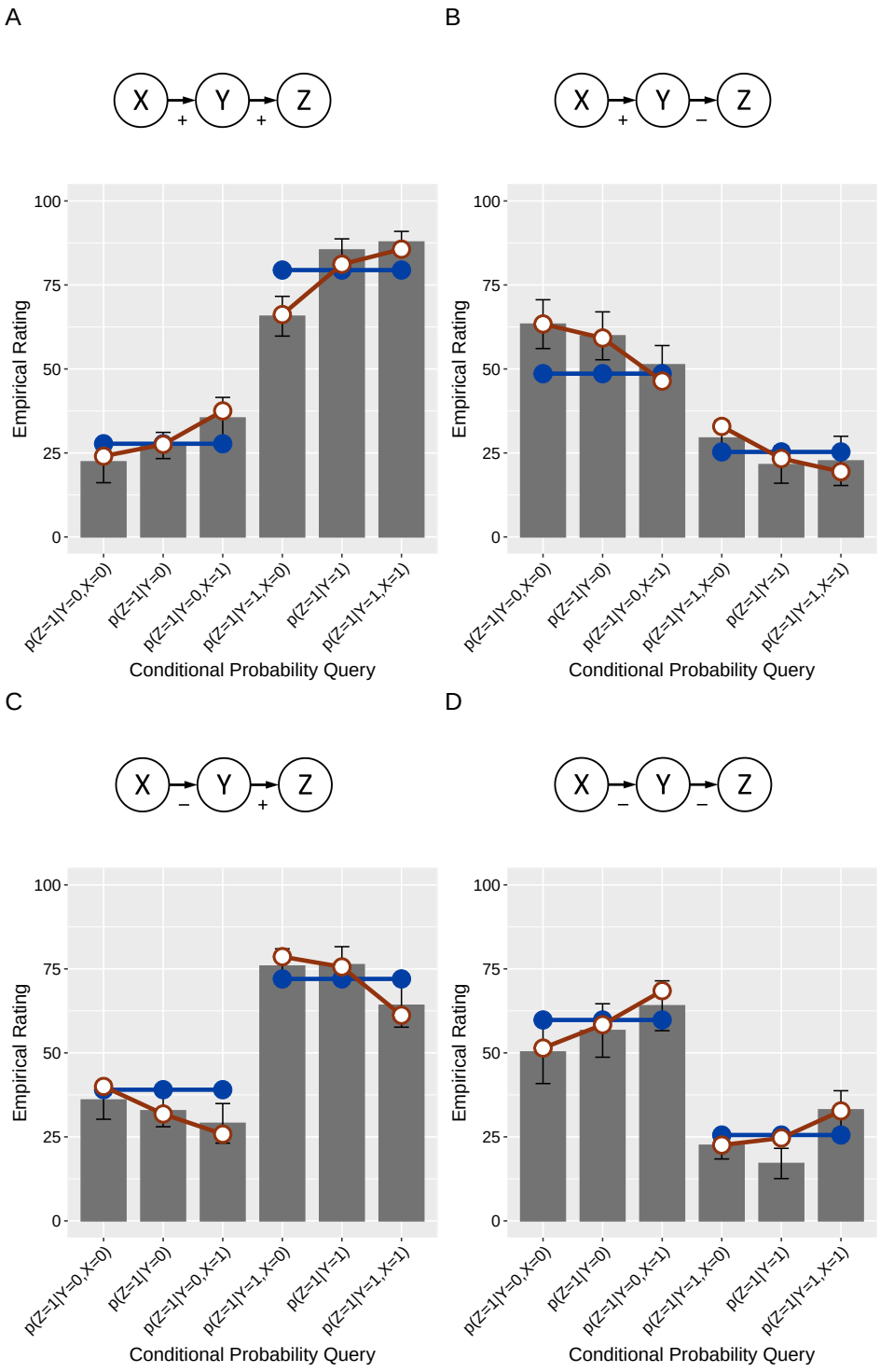

Figure 9: Noisy logical fits to tests items in Experiment 2. A. Generative-generative condition. B. Generative-inhibitory condition. C. Inhibitory-generative condition. D. Inhibitory-inhibitory condition. The blue and red lines are the fits of the normative model and the mutation sampler, respectively. Error bars are 95% confidence intervals.

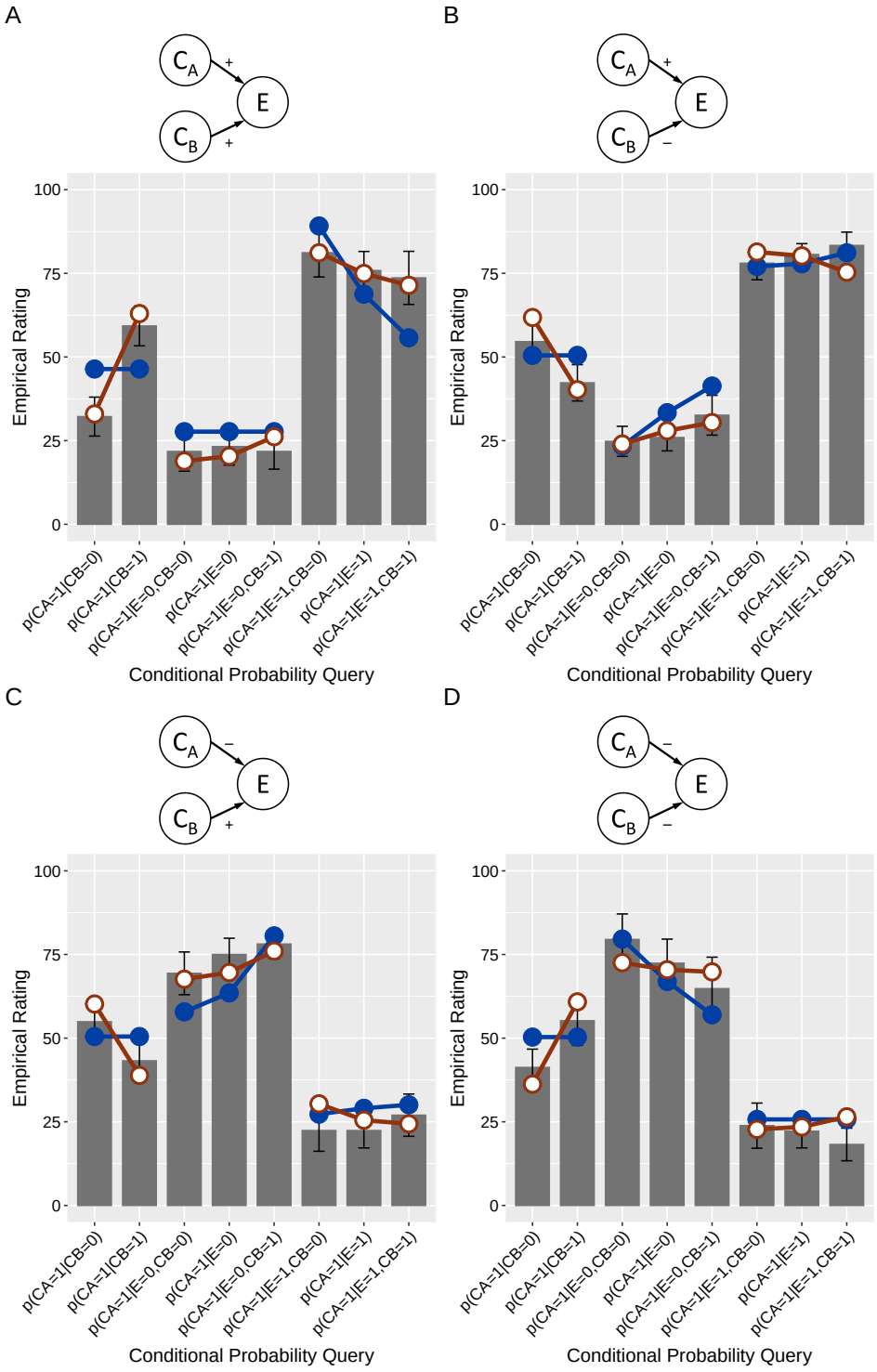

Figure 10: Noisy logical fits to tests items in Experiment 3. A. Generative-generative condition. B. Generative-inhibitory condition. C. Inhibitory-generative condition. D. Inhibitory-inhibitory condition. The blue and red lines are the fits of the normative model and the mutation sampler, respectively. Error bars are 95% confidence intervals.

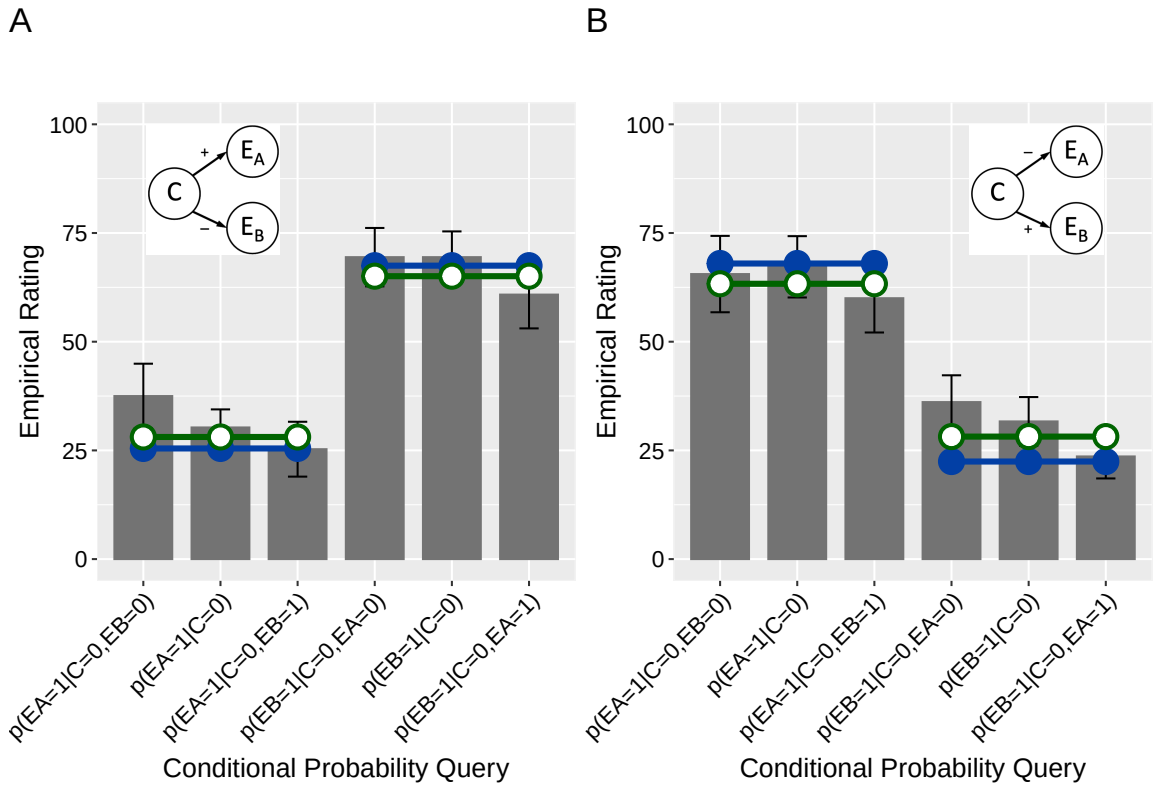

Figure 11: Fits of the normative (non-mutation sampler) model assuming logistic (blue lines, closed plot points) versus noisy-logical (green lines, open plot points) generating functions to the generative-inhibitory (Panel A) and inhibitory-generative (B) conditions of Experiment 1. The noisy-logical model has two free parameters representing the exogenous causes of E_A and E_B.

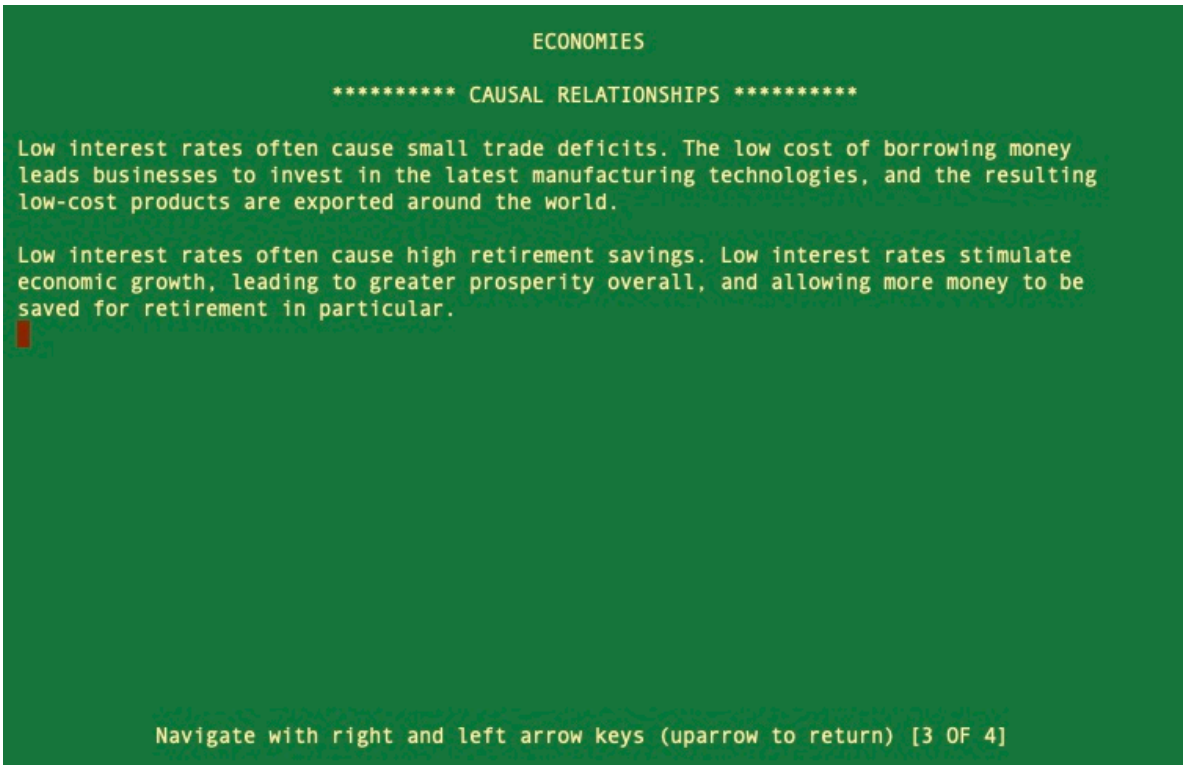

Figure 12: Example of the causal relationships presented to subjects.

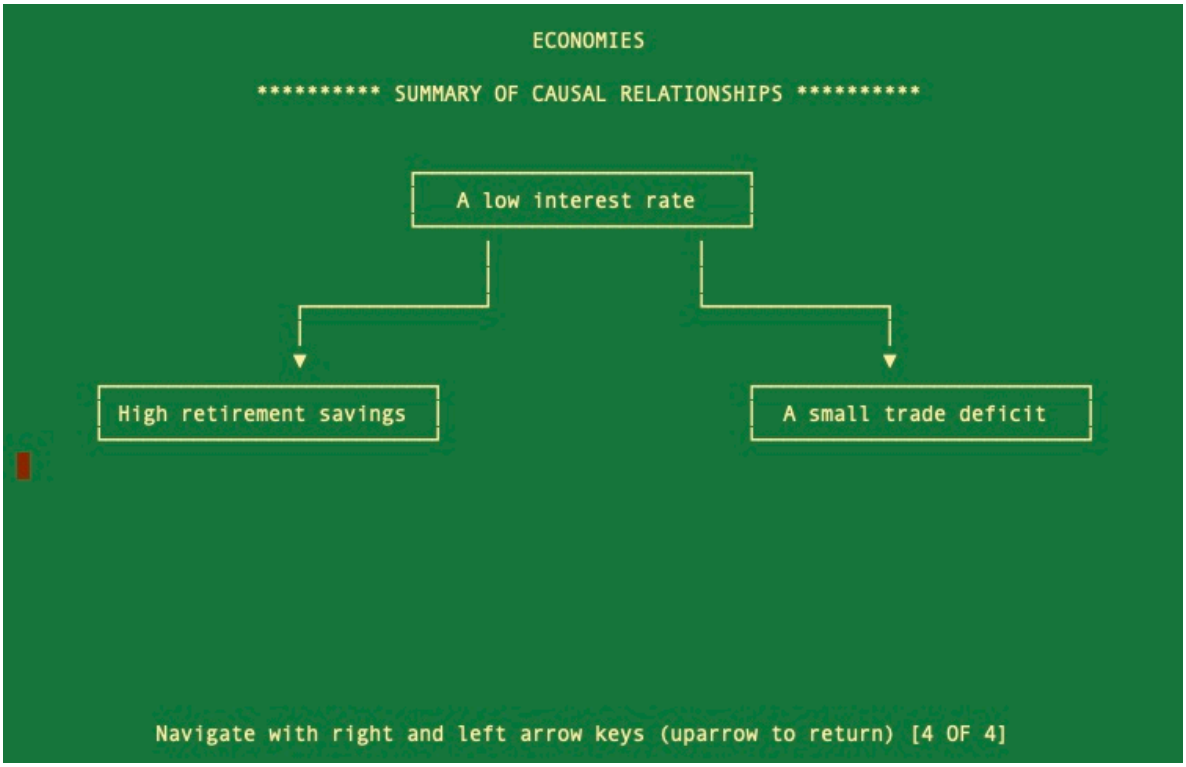

Figure 13: Example of the causal diagram presented to subjects.

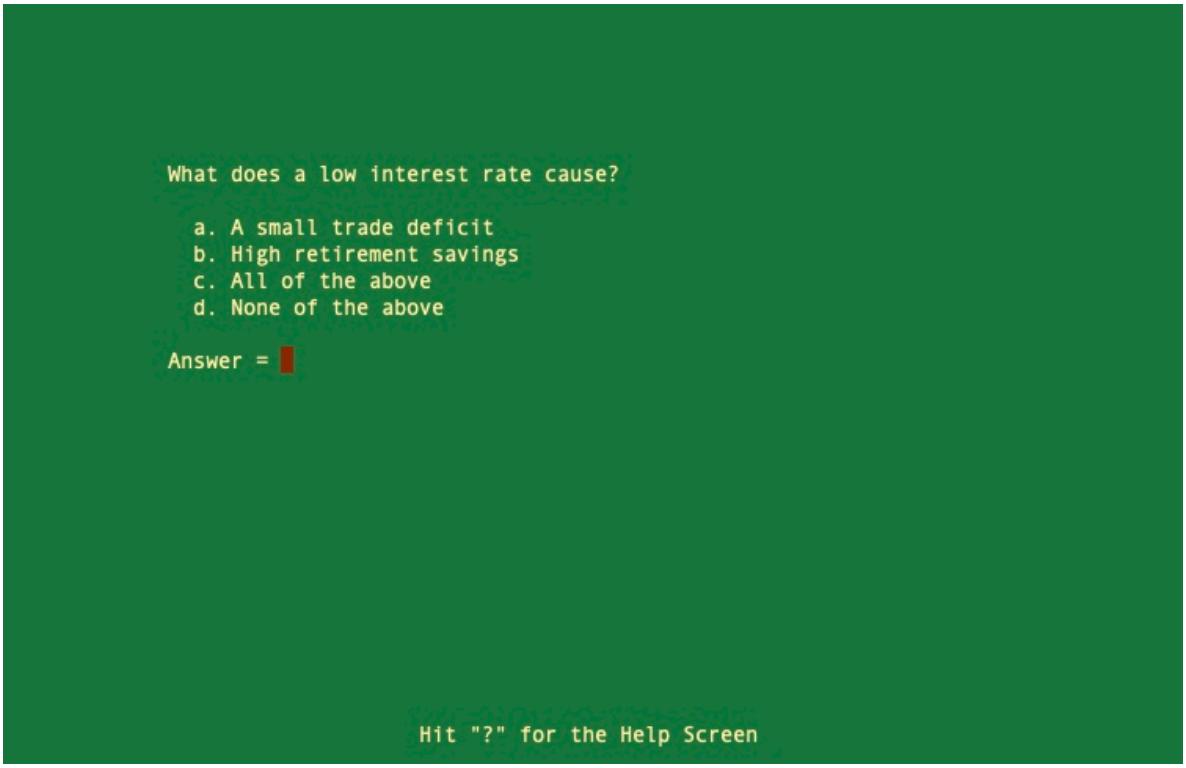

Figure 14: Example of a multiple-choice question presented to subjects.

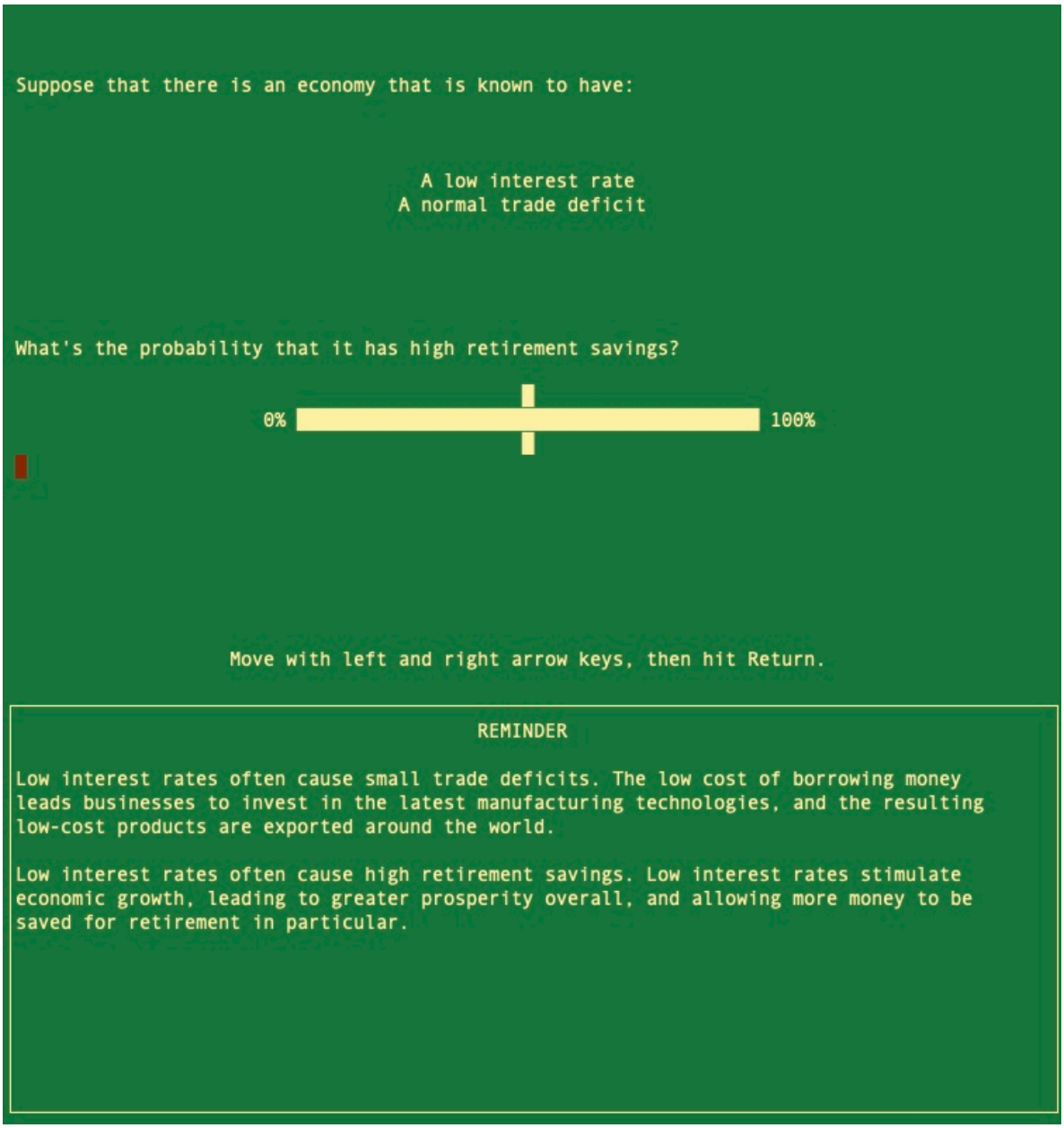

Figure 15: Example of a multiple-choice question presented to subjects.