# Causal Schema-Based Inductive Reasoning

**Oren Griffiths (ogriffiths@psy.unsw.edu.au)**
**Ralf Mayrhofer (rmayrho@uni.goettingen.de)**
**Jonas Nagel (nagel.jonas@gmail.com)**
**Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)**
Department of Psychology, University of Göttingen, Germany

**Abstract**

Inductive reasoning allows us to go beyond the target hypothesis and capitalize on prior knowledge. Past research has shown that both similarity relations and specific causal knowledge affect the inductive plausibility of hypotheses. The present experiment goes one step further by investigating the role of abstract causal schemas about main effects and interactions. We were interested in exploring whether the functional form of a causal schema influences our inductions even when no more specific causal knowledge is available. Our experiment shows that reasoners have different prior beliefs about the likelihood of main-effect versus interactive schemas, and rationally combine these prior beliefs with new evidence in a way that can be modeled as Bayesian belief updating. This finding casts doubt on theories which ignore the important role of priors in inductions involving causal schemas.

**Keywords:** Inductive Reasoning; Causal schemas; Causal Interactions; Rational model; Bayesian inference; Computational Modeling

## Introduction

Inductive reasoning is ubiquitous. In associative learning we infer from a set of learning trials to a general regularity which probably, or so we assume, also applies in the future. In causal learning we use a sample of observations and go beyond the information given to induce general causal laws, which underlie our predictions, diagnoses, and our action plans (see Waldmann, Hagmayer, & Blaisdell, 2006). Inductions not only occur at the level of exemplars but also on the level of prior knowledge about hypotheses. For example, knowing that dogs have hearts allows us to give an informed guess about the probable validity of the hypothesis that wolves have hearts, as well. The interconnectedness of our knowledge is a powerful tool to quickly gain knowledge by mining our prior knowledge to yield inductive biases.

Although inductive inferences have for a long time been studied in learning, inductions between hypotheses is a fairly recent research goal (see Feeney & Heit, 2007, for an overview). Many early studies have focused on categories as the basis for inductive inference. For example, the inference from dogs to wolves mentioned in the last paragraph is probably driven by the similarity of the categories dogs and wolves. In their seminal article Osherson, Smith, Wilkie, Lopez, and Shafir (1990) proposed a model ("similarity coverage model") that implies that the strength of an argument depends on both (i) the similarity between the premise categories and the conclusion categories, and (ii) the extent to which the premise categories provide good coverage of the category to which the conclusion statement refers.

Competing models of category-based induction have been proposed. For example, Heit (1998) has proposed a *Bayesian* model of inductive reasoning. According to this model, evaluating an inductive argument is conceived of as learning about a property. The key assumption is that the inductive generalization of a novel property from one category to a second category is sensitive to prior knowledge about other properties the two categories share. Thus, we tend to generalize properties from dogs to wolves because they share a lot of properties so that they probably also share this novel property. In contrast, we are more reluctant to generalize from dogs to parrots because we believe them to share few properties.

## The Role of Causal Knowledge

Similarity-based models have their limits. They typically cover cases in which blank predicates that do not invoke prior knowledge are used. However, once meaningful predicates are used it becomes clear that similarity is not the only factor that drives inductive inferences. A number of studies have demonstrated that causal relations between categories and features become important with meaningful predicates. For example, Heit and Rubinstein (1994) have shown in an early study that a behavioral property (e.g., travels in zig-zag path) was more strongly generalized from tuna to whale than from bears to whales. In contrast, a biological property (e.g., has a two-chambered liver) is generalized more from bears to whales instead. A plausible explanation is that in these inductions reasoners consider the kinds of common causal mechanisms that could generate these properties. Tuna and whales may share a common behavioral property because they live in a similar ecology; whereas common biological properties are more likely to arise in organisms that are taxonomically similar (see also Sloman, 1994). Another example of the role of causal knowledge is the finding that undergraduates are more likely to infer that "monkeys have enzyme X" from the premise "bananas have enzyme X" than from "mice have enzyme X," even though mice are more similar to monkeys than to bananas (Medin, Coley, Storms, & Hayes, 2003). Apparently prior knowledge about the possible causal relation between eating bananas and transferring enzymes is activated in this case. Based on a variant of causal-model theory Rehder (2007) has proposed a theory which treats

generalizations as causal reasoning. According to this theory we assess the likelihood that a novel feature applies to a new category on the basis of our beliefs about the causal relations that connect that feature to the category. Finally, a Bayesian model integrating a large number of findings comes from Tenenbaum, Kemp, and Shafto (2007). In this model a Bayesian inference engine is coupled with theory-based structured priors. Depending on the induction task these priors can take the form of taxonomies or can embody causal knowledge (e.g., about food chains).

## Causal Schemas

The examples discussed in the previous section illustrate the role of causal knowledge, which can be specific (monkeys ingest the ingredients of food) or more abstract (biological kinds share essences which give rise to common properties). Schematic causal knowledge may be even more abstract. Kelley (1972) has proposed schemas for multiple necessary or multiple sufficient causes which are domain-general but nevertheless aid inferences (e.g., discounting). These schemas specify prior assumptions about the way multiple causes collaborate when jointly generating or preventing an effect. Waldmann (2007) has shown that different domains trigger different causal integration schemas. For example, when the causal effect is the likeability of objects participants tended to average the causal influences. With other effects they preferred to add them. In the causal Bayes net literature causal schemas have been discussed under the label functional form. Tenenbaum and Griffiths (2003) argued that domain knowledge not only constrains causal structure but also functional forms that specify the relation between multiple causes and effects. As an example, they modeled Sobel, Tenenbaum and Gopnik's (2004) findings using a Bayes net in which prior knowledge about causal schemas is integrated. Sobel et al. had argued that children's performance can be best explained if it is assumed that they enter the task with the prior assumption that the individual causes do not interact (i.e., "noisy-or" rule). Novick and Cheng (2004) have also analyzed the question of how main effects (additive integration) can be differentiated from causal interactions in the framework of power PC theory. Within this model, main effects are considered the default causal schema, and interactions are represented as deviations from the default.

## Patterning Interactions

Interestingly, there is a long history of studying interactions in associative learning, although they are typically referred to as "patterning" interactions in this literature. Positive patterning (PP) refers to learning a situation in which two cues (e.g. A & B), when presented individually, are not paired with the outcome (A- and B- trials), but when presented together they are paired with the outcome (AB+ trials). In contrast, negative patterning (NP) refers to a scenario in which cues A & B, when presented alone, are paired with the outcome (A+ and B+ trials), but when presented together they are not paired with the outcome (AB- trials). Shanks and Darby (1998) found that people can learn both of these interactions (PP and NP) concurrently, and can form the appropriate abstract schematic representations. Consequently, after being repeatedly shown two cues (A and B) interacting, they can infer that the same interaction will occur between two other cues (C and D) which had not previously been shown together. For example, participants that underwent NP training with cues A and B, and were then shown C+ and D+ trials, could infer that the novel compound CD would not be followed by the outcome. That is, participants constructed (or selected) a causal schema to describe their observations of cues A and B, and then inductively inferred that this schema may apply to two novel cues, C and D.

Kemp, Goodman, and Tenenbaum (2007) have recently proposed a Bayesian model which can account for Shanks and Darby's (1998) data. The model achieves this by learning causal schemas; that is, the model monitors the co-occurrences of cues and outcomes, and groups together cues that behave in a similar manner in training. In the NP case, this model groups together cues that co-occur with the outcome in isolation, but do not co-occur with the outcome when paired with another cue of the same kind. Importantly, the model can use these cue groupings to generate predictions about novel cue-combinations at test, and thus solve the [C+, D+, CD?] test cases.

In line with the results of Shanks and Darby (1998) the model assumes that if a participant is shown two cues interacting (A and B), then that participant will generally consider this interaction as informative as to whether an interaction will occur between two further cues (C and D). However, the informativeness or inductive support of an instance of an interaction schema also depends upon the structure of participants' prior beliefs regarding the likelihood of the various possible causal schemas. Although Shanks and Darby's learners seemed to frequently transfer the observed patterning rules to new features, we believe that this may not be a general reasoning pattern but rather be due to the demand characteristics of the experiment. When confronted with a large set of cases of interactions, participants may have reasoned that what is expected from them is to abstract and transfer a rule. The experiments show that participants are able to do this, but this may not be what they would do under more natural circumstances.

There is some research suggesting that participants' initial beliefs about causal schemas are biased against patterning interactions. For example, Novick and Cheng (2004) have suggested that people treat main-effect causal schemas (e.g. A+, B+, AB+) as a kind of default assumption, and thus suggest that the initial likelihood of main-effect schemas may be high. Consistent with this assumption, studies of patterning have shown that participants often demonstrate difficulty applying the patterning schemas on the novel test cases (Kehoe, 1988). Moreover, studies about cue integration in different types of judgments generally reveal additive, linear integration (Dawes & Corrigan, 1974). Thus,

it is reasonable to expect an initial bias against interactive patterns, especially disordinal ones as in the NP case.

Learning tasks in which participants are repeatedly presented with many individual cases, such as that used by Shanks and Darby (1998), may be less suited for revealing such biases than inductive reasoning tasks that rely upon general hypotheses. Imagine, for example, somebody informs you that two drugs A and B which each cure a disease generally cancel each other's effect when taken together. Would we really expect two arbitrary different drugs C and D behave similarly?

In the present research we have decided to approach the question of how people transfer knowledge about causal schemas using a reasoning task in which participants are presented with hypotheses rather than individual cases. Our main goal was (i) to investigate prior beliefs about novel hypotheses embodying different types of schemas (as in the drug example), and (ii) to study how hypothetical knowledge of the truth of a similar hypothesis conforming to the same schema will influence these beliefs.

## Schema-based Priors and Belief Updating

Participants' prior schematic beliefs regarding main effects and interactions have not previously been assessed directly. Thus, the first goal of the present research is to examine participants' prior beliefs regarding causal schemas by explicitly asking them to rate how plausible they believe NP and PP interactions to be, compared to their corresponding main effects (explained below). In the NP case, i.e., A and B are paired with the effect when independently shown, there are two hypotheses about the conclusion that can be drawn with respect to the effect's occurrence when A and B are presented together

$$H_{NP}: \quad (A+, B+) \Longrightarrow (AB-)$$
$$H_{ME+}: \quad (A+, B+) \Longrightarrow (AB+)$$

The first hypothesis refers to NP: the compound doesn't bring about the effect, although both causes individually do. The second hypothesis refers to the corresponding main effect (which we call ME+). Obviously, these two hypotheses are complementary, i.e. $P(H_{NP}) + P(H_{ME+}) = 1$, because the effect can either occur or not occur.

The same can be derived for the PP case, i.e. A and B are *not* paired with the effect when presented alone:

$$H_{PP}: \quad (A-, B-) \Longrightarrow (AB+)$$
$$H_{ME-}: \quad (A-, B-) \Longrightarrow (AB-)$$

Accordingly, the first hypothesis describes PP: the compound does bring about the effect whereas both causes separately do not. The second hypothesis is the complimentary main effect (which we call ME-).

In these terms our first prediction is that participants will consider patterning schemas to be less plausible than the corresponding control schemas (i.e., main effects), this can be expressed as $P(H_{NP}) < P(H_{ME+})$ as well as $P(H_{PP}) < P(H_{ME-})$.

A second question addresses the extent to which participants change their belief in an initial hypotheses $H_i$ in response to a conforming instance $D_i$, e.g., when knowing that the conclusion $(C+, D+) \Longrightarrow (CD-)$ is true for novel cues C and D from the same domain as A and B in the case of a PP hypothesis. Bayes' rule suggests that the informativeness of such an instance depends upon the structure of participants' prior beliefs regarding the possible hypotheses and the likelihood of the instance given the hypotheses:

$$P(H_i|D_i) \propto P(D_i|H_i)P(H_i)$$

For the sake of simplicity, we assume that the likelihood of an instance $D_i$ given main-effect or patterning hypotheses is a function of its similarity to the instance addressed by the hypotheses, and is therefore independent of the type of the schema.[1] For instances conforming with the schema in the hypothesis this assumption can be represented by some fixed number larger than 0.5 (i.e., the instance is informative) but less than 1 (i.e., the inference from the instance to the hypothesis, which is formulated with respect to another pair of cues, is tainted with uncertainty).

If so, then the lower the initial belief in a hypothesis, the larger the change in belief will be upon encountering evidence consistent with that belief. With respect to the present experiment, an instance of an unlikely patterning schema ought to be more informative (i.e., provide more inductive support) than a confirmatory instance of a likely main-effect schema. Thus, prediction (ii) is that participants will change their beliefs regarding patterning schemas more in response to an observed instance of a patterning interaction, than they will for the corresponding main-effect schemas:

$$P(H_{NP}|D_{NP}) - P(H_{NP}) > P(H_{ME+}|D_{ME+}) - P(H_{ME+})$$
$$P(H_{PP}|D_{PP}) - P(H_{PP}) > P(H_{ME-}|D_{ME-}) - P(H_{ME-})$$

Finally, this framework provides a third prediction regarding participants' posterior belief in main-effect and patterning schemas. Although a greater increase in belief is anticipated for the patterning schemas in response to a patterning instance, than for the main-effect schemas in response to a main-effect instance, the posterior belief for the patterning schemas cannot exceed that of the main-effect schemas as long as the likelihood of a conforming instance is independent of the current hypothesis:[2]

$$P(H_{NP}|D_{NP}) < P(H_{ME+}|D_{ME+})$$
$$P(H_{PP}|D_{PP}) < P(H_{ME-}|D_{ME-})$$

Thus, it is predicted that after being shown a confirming instance, participants' belief ratings for the patterning

[1] So, $P(D_i|H_i) = P(D_j|H_j) \ \forall\, i, j \in \{NP, ME+, PP, ME-\}$

[2] This easily follows from our first prediction by backward application of Bayes' rule and $P(D_i|H_i) = P(D_j|H_j)$, e.g.: $P(H_{ME+}) > P(H_{NP}) \Rightarrow P(D_{ME+}|H_{ME+})P(H_{ME+}) > P(D_{NP}|H_{NP})P(H_{NP}) \Rightarrow P(H_{ME+}|D_{ME+}) > P(H_{NP}|D_{NP})$

conclusion statements will be lower than their belief ratings for the main-effect conclusion statements.

Finally, the domains of the causal scenarios presented to participants were manipulated. Participants may have different prior beliefs regarding how plausible particular causal schemas are in certain domains (e.g. interactions between chemical substances may be more plausible than interactions between social causes; see Waldmann, 2007; Wattenmaker, 1995). The present research also sought to provide exploratory data on this issue.

## Method

**Participants.** Thirty two undergraduate students, mostly from the University of Göttingen, participated. They were either given course credit or €3 for participating.

**Design.** The scenarios given to participants varied with respect to two factors (each with four levels). The first factor was the type of schema depicted: PP, ME-, NP or ME+. Table 1 summarizes each of these interactions (the distinction between Questions 1 and 2 is addressed in the Procedure). The second factor was the domain of the causal relationship: psychological, biological, physical, or chemical.

Table 1. Design of Experiment.

| | *Cue Interaction* | | | |
|---|---|---|---|---|
| | PP | ME- | NP | ME+ |
| *First Question* | A- | A- | A+ | A+ |
| | B- | B- | B+ | B+ |
| | AB+ | AB- | AB- | AB+ |
| *Second Question* | C- | C- | C+ | C+ |
| | D- | D- | D+ | D+ |
| | CD+ | CD- | CD- | CD+ |
| | A- | A- | A+ | A+ |
| | B- | B- | B+ | B+ |
| | AB+ | AB- | AB- | AB+ |

*Note*: Letters A – D represent causes, and symbols + and – indicate statements in which the cause either produce the effect or do not, respectively. Statements above the dashed lines are premises, and the statements below the dashed lines are conclusions.

A partially confounded within-by-between (Latin Squares) design was used. This means that each participant experienced 8 scenarios (two of each schema/domain), rather than all 16 possible combinations. Sixteen versions of an eight-scenario questionnaire were constructed in order to counterbalance the assignment of question domain to causal schema. Thus, across participants, each relationship (e.g., PP, ME-) occurred equally frequently in each domain (e.g., psychological, physical). The question order was also counterbalanced between subjects, so that an equal number of participants saw (i) an NP relationship before a PP relationship as vice versa, and saw (ii) a main-effect relationship before a patterning relationship or vice versa. A similar counterbalancing methodology was applied to the scenario domains, but this was performed orthogonally to that for causal relationship so as to not produce any systematic relationship between the domain of each scenario, and the causal schema depicted. Finally, the assignment of cues to the premise or conclusion statements was counterbalanced between individuals. For example, participants were equally often required to infer from a premise involving gold and copper to a conclusion involving silver and lead, as from a premise involving silver and lead to a conclusion involving gold and copper.

**Materials and Procedure.** Each participant completed a questionnaire that consisted of a short instruction section, an explanatory example scenario, and then 8 causal scenarios. Each questionnaire was 20 pages long and was presented in German. The instructions informed participants that in each scenario they would first be given two premise statements, and would then have to rate to what extent they believed a given conclusion statement to be true. They were told to assume that the premise statements were true when judging the conclusion statement. Finally, participants were told that they would rate each conclusion statement twice, but that they would be given more information between the first and second rating occasions.

All 8 scenarios each described four cues (A – D) that either did or did not cause the outcome. All scenarios involved unfamiliar causal statements, whereby participants were told about real items causing fictitious effects. Fictitious effects were chosen to ensure that participants did not rely on specific prior knowledge about the causal relations. The scenarios each followed the same two-question format. Participants were first shown two premise statements, in each of which one cue A or B was shown to individually cause the outcome, or not (e.g., Facts A and B alone in Table 2). Participants were then given a causal statement involving compound AB either causing, or not causing, the outcome. They were asked to rate how true they believed this conclusion statement to be (referred to as Question 1)(e.g., Conclusion in Table 2). To do this, participants were provided with an 11-point scale, labeled "definitely false" at the left-hand end (0) and "definitely true" at the right-hand end (10).

Table 2. A sample NP trial (Question 2)

| | |
|---|---|
| *Fact 1*: | Exposing compound 3X8 to nitrogen gas causes the compound to become brittle. |
| *Fact 2*: | Exposing compound 3X8 to neon gas causes the compound to become brittle. |
| *Fact 3*: | Exposing compound 3X8 to both nitrogen and neon gases do not cause the compound to become brittle. |
| *Fact A*: | Exposing compound 3X8 to oxygen gas causes the compound to become brittle. |
| *Fact B*: | Exposing compound 3X8 to argon gas causes the compound to become brittle. |
| *Conclusion*: | Exposing compound 3X8 to both oxygen and argon gases do not cause the compound to |

become brittle.

After answering Question 1, participants turned the page and were shown Question 2, in which some further information about that same scenario was presented. The first two facts (Facts A and B) and the conclusion statement were repeated. Participants were additionally provided with three new premise statements, Facts 1 – 3, positioned above the initial two premise statements (as shown in Table 2). The three new facts described the causal relationship between the outcome and two new cues from the same domain (C and D) when those cues were presented individually (Facts 1 and 2) and when they were presented in compound (Fact 3). Facts 1 – 3 always displayed the same causal main effect or interaction as Facts A, B, and Conclusion (NP, PP, ME+, ME-). That is, Facts 1 – 3 together constituted an example of the same causal schema (using different cues) that participants were required to assess in the conclusion statement. In Question 2, participants rated the same conclusion statement as in Question 1, and used the same scale as was used in Question 1. Participants then proceeded directly to the next scenario. This process was repeated until all 8 scenarios were complete. Participants were not allowed to return to any previous questions.

## Results

Figure 1 depicts participants' mean responses for Questions 1 and 2, for each of the four types of causal relationships (NP, PP, ME+, ME-). Type I error (α) was controlled at .05, andthree planned contrasts were tested using a two-way (schema-type and question number) ANOVA. .

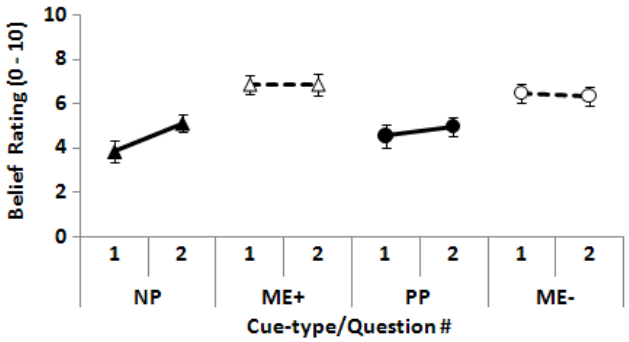


Figure 1. Mean conclusion belief ratings (±SEM) for patterning and main-effect schemas. Triangles indicate responses to NP and ME+ questions, and circles represent PP and ME- questions. Solid lines represent belief-updating for patterning interactions, and dashed lines represent belief updating for main effects.

As anticipated, participants' judgments on Question 1 for the patterning interactions were lower than for the main-effect conditions, $F(0.05/3,1,31) = 26.25$, $p < .05$. Similarly, participants' judgments on Question 2 were lower for the patterning interactions (PP and NP) than for the main-effect conditions (ME- and ME+), $F(0.05/3, 1,31) = 11.31$, $p < .05$. A significant interaction between the two factors revealed that the difference between participants' ratings on Questions 1 and 2 was significantly larger for the causal interaction conditions (NP, PP) than for the causal main-effect conditions (ME-, ME+), $F(0.05/3, 1,31) = 13.66$, $p < .05$. That is, participants changed their belief between Questions 1 and 2 more on the trials that depicted a causal interaction than on those that depicted a causal main effect.

Due to the low number of participants in each relevant counterbalancing group ($N = 8$) conclusions about individual domains are only tentative. Nevertheless, this data were analyzed using two-tailed *t*-tests. In each domain participants initially (on Question 1) rated the patterning interactions (averaged across NP and PP) as less plausible than the main-effect scenarios, but these differences were not significant, maximum $t(7) = 2.06$.

## General Discussion

In the experiment we found that prior to being presented with data, patterning interactions were considered to be less plausible than their respective main-effect schemas both before (Question 1) and after (Question 2) a different example of that causal schema from the same domain was presented. Moreover, the addition of an example of two cues demonstrating an NP or PP interaction produced a greater increase in belief than an example of two cues demonstrating a main-effect relationship. In other words, an instance of a patterning interaction provides more inductive support than an instance of a main-effect schema, although the posterior beliefs about interactive patterns still proved well below the belief for main effects. This pattern is in line with rational Bayesian reasoning. Additional evidence should have more impact on hypotheses that are initially deemed implausible than on plausible ones, which are already near ceiling. Our results generally demonstrate the role of prior beliefs about abstract causal schemas in the inductive evaluation of hypotheses.

Participants considered interaction schemas to be generally less plausible than main effects in all four domains (intuitive psychology, biology, chemistry and physics), and no differences were seen between these domains. However, due to the relative insensitivity of our small sample ($N = 8$ in some analyses) potential domain effects could not reveal themselves if they exist. It would be interesting to further explore whether different domains yield differences in biases regarding abstract causal schemas (cf. Wattenmaker, 1995).

In the present research we have empirically tested the hypothesis that interaction schemas are considered rare compared to main effect schemas. It would be interesting to investigate where this bias comes from. One possibility is that interactions are less frequent in our environments than linear relations, and that our beliefs about causal schemas simply reflect differential frequency. We believe that this account oversimplifies the situation. Our causal representations are not only determined by the structure of

the world but also by the categories we acquire in our phylo- and ontogeny to represent the world. For example, Clark and Thornton (1997) have shown that interactions (e.g., XOR structures) can be recoded as linear relations with the right choice of features and categories (see also Waldmann & Hagmayer, 2006).

Thus, a more plausible hypothesis is that people have a general tendency to favor categories and hypotheses yielding simple causal explanations rather than complex, interactive ones. Interactions force us to additionally represent co-factors when considering an individual cause which is more taxing for our information processing capacity in reasoning and learning than context independent linear relations. Thus, phylogenetically a preference may have evolved to induce categories that preferentially yield linear relations between multiple causes. Moreover, we may have a bias towards representing ordinal interactions as linear, even when we slightly distort reality. Dawes and Corrigan (1974) have shown that ordinal interactions can robustly be approximated by linear relations with no substantial loss. Thus, our bias against interactions may be a joint product of the world and our mind.

## Acknowledgements

We wish to thank Mira Holzer for assistance with data collection, and York Hagmayer for helpful comments on the project. This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (DFG, Wa 621/20-1) and of the Deutscher Akademischer Austauschdienst (DAAD, A/08/94580). The first author (O.G.) is now at the University of New South Wales, Sydney, Australia.

## References

Clark, A., & Thornton, C. (1997). Trading spaces: Computation, representation and the limit of uninformed learning. *Behavioral & Brain Sciences*, *20*, 57-90.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95-106.

Feeney, A., & Heit, E. (2007). *Inductive reasoning: Experimental, developmental, and computational approaches.* New York, NY: Cambridge University Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition*. Oxford: Oxford University Press.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 411-422.

Kehoe, E. J. (1988). A layered network model of associative learning: Learning to learn and configuration. *Psychological Review*, *95*, 411-433.

Kelley, H. H. (1972). *Causal schemata and the attribution process*. New York: General Learning Press.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemas. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (pp. 64-70).

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*, 517-532.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*, 455-485.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category based induction. *Psychological Review*, *97*, 185-200.

Rehder, B. (2007). Property generalization as causal reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches*. New York: Cambridge University Press.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 405-415.

Sloman, S. A. (1994). When explanations compete: the role of explanatory coherence on judgments of likelihood. *Cognition*, *52*, 1-21.

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

Tenenbaum, J. B. Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney & E. Heit, (Eds.), *Inductive reasoning*. New York, NY: Cambridge University Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science*, *31*, 233-256.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*, 27-58.

Waldmann, M. R., Hagmayer, Y, & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*, 307-311.

Wattenmaker, W. D. (1995). Knowledge structures and linear separability: Integrating information in object and social categorization. *Cognitive Psychology*, *28*, 274-328.