The Assumption of Class-Conditional Independence in Category Learning

Jana Jarecki (jarecki@mpib-berlin.mpg.de)*

Björn Meder (meder@mpib-berlin.mpg.de)*

Jonathan D. Nelson (nelson@mpib-berlin.mpg.de)*

*Center for Adaptive Cognition and Behavior (ABC),
Max Planck Institute for Human Development, Lentzeallee 94
14195 Berlin, Germany

Abstract

This paper investigates the role of the assumption of class-conditional independence of object features in human classification learning. This assumption holds that object feature values are statistically independent of each other, given knowledge of the object's true category. Treating features as class-conditionally independent can in many situations substantially facilitate learning and categorization even if the assumption is not perfectly true. Using optimal experimental design principles, we designed a task to test whether people have this default assumption when learning to categorize. Results provide some supporting evidence, although the data are mixed. What is clear is that classification behavior adapts to the structure of the environment: a category structure that is unlearnable under the assumption of class-conditional independence is learned by all participants.

Keywords: Multiple-cue classification learning; class-conditional independence; naïve Bayes; causal Markov condition

Introduction

Categorization is fundamental for cognition. Grouping together objects or events helps us to efficiently encode environmental patterns, make inferences about unobserved properties of novel instances, and make decisions. Without categorization we could not see the woods for the trees.

Despite the ease with which we form categories and use them to make inferences or judgments, from a computational perspective categorization is a challenging problem. For instance, different diseases can cause similar symptoms, entailing that diagnostic inferences are often only probabilistic. Patients may have new symptom combinations and still require a diagnosis. Depending on the specific assumptions the physician makes about the relationship between the diseases and symptoms, a physician could justifiably make very different inferences about the diseases.

In the present paper, we investigate the role of the possible assumption of *class-conditional independence* of features in category learning. Class-conditional independence holds if the features of the category members are statistically independent given the true class. This assumption can facilitate classification and learning of category structures. The concept of class-conditional independence underlies the naïve Bayes classifier in machine learning (Domingos & Pazzani, 1997), and is also a key assumption in some psychological classification models (e.g., Fried & Holyoak, 1984; Anderson, 1991). It is related to ideas of channel separability in sensory perception (Movellan & McClelland, 2001). Similar ideas are found in Reichenbach's (1956) *common-cause principle* in the philosophy of science and in causal modeling (Spirtes, Glymour, & Scheines, 1993; Pearl, 2000).

Both the philosophical and psychological literature make claims about the normative bases of the assumption of class-conditional-independence of features. Our focus here is not on the general normativity or nonnormativity of that assumption, but on whether the assumption of class-conditional independence may (perhaps tacitly) underlie people's inferences in learning and multiple-cue categorization tasks. We think of this assumption as one of many possible default (heuristic or meta-heuristic) assumptions that, if close enough to an environment's actual structure, may facilitate learning and inferences.

The Psychology of Conditional Independence

Some psychological models of categorization incorporate assumptions of class-conditional independence, such as the *category density model* (Fried & Holyoak, 1984) or Anderson's (1991) *rational model of categorization*. Both models treat features of instances as class-conditionally independent to make inferences about category membership or unobserved item properties.

Other research has focused more directly on the role of conditional independence assumptions in human reasoning. For instance, a key assumption in many formal causal modeling approaches (e.g., Pearl, 2000; Spirtes et al., 1993) is the so-called *causal Markov condition*, which assumes that a variable in a causal network is independent of all other variables (except for its causal descendants), conditional on its direct causes. As this assumption facilitates probabilistic inferences across complex causal networks it was suggested that people's causal inferences could also comply with this conditional independence assumption.

Von Sydow, Meder, and Hagmayer (2009) investigated reasoning about causal chains and found that subjects' inferences indicated a use of conditional independence assumptions, even if the learning data suggested otherwise.[1] Other research, however, found violations of the causal Markov condition (Rehder & Burnett, 2005). Asked to infer the prob-

[1] For instance, applying the causal Markov condition to a causal chain $X \rightarrow Y \rightarrow Z$ entails that Z is independent of X given Y (e.g., $P(z|y,x) = P(z|y,\neg x)$).

ability for one effect when knowing the common cause of several effects, people's judgments were influenced by the status of the other effects rather than treating all effects as independent of each other given the cause. One explanation for this "nonindependence effect" (Rehder & Burnett, 2005) is that it might be due to subjective explanations that disable all causal links between the cause and effects at once (Walsh & Sloman, 2007). Other researchers have argued that these Markov violations do not indicate flawed human reasoning, but reflect the use of abstract causal knowledge that is sensitive to contextual information (Mayrhofer, Hagmayer, & Waldmann, 2010).

Research Questions

Should the assumption of class-conditional feature independence be used in classification learning? Do people use that assumption to guide learning about the structure of a novel environment? We extend previous research fourfold: (1) We use optimal experimental design principles (Myung & Pitt, 2009; Nelson, 2005) to *explicitly* address the assumption in classification, (2) we are interested in categorization *learning* as opposed to causal reasoning, (3) we investigate how people's experience with a new environment shapes their classification *behavior*, whereas many previous studies have measured explicit numerical probability judgments. (4) We use an *experience-based* research paradigm, whereas previous studies used numerical (Rehder & Burnett, 2005) or verbal (Mayrhofer et al., 2010) formats. Personal experience of events has been shown to result in different behavior and learning than word- or number-based presentation of probabilities (Hertwig, Barron, Weber, & Erev, 2004; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Before describing the task we designed, let us turn to the normative question of class-conditional independence in classification.

Class-Conditional Independence in Classification

Categorization entails assigning an object to a class. Let F denote an object consisting of a vector of feature values $\boldsymbol{f}$, and let C denote a random variable whose values are the possible classes $c_1, \ldots, c_n$. The posterior probability of the class given the observed feature values, $P(\text{class} \mid \text{features})$, can be inferred using Bayes' rule:

$$P(C = c \mid F = \boldsymbol{f}) = \frac{P(F = \boldsymbol{f} \mid C = c)P(C = c)}{P(F = \boldsymbol{f})} \qquad (1)$$

where $P(F = \boldsymbol{f} \mid C = c)$ denotes the likelihood of feature value vector $\boldsymbol{f}$ given class c, $P(C = c)$ is the prior probability of the class, and $P(F = \boldsymbol{f})$ is the occurrence probability of the feature configuration. An important question is how we estimate the relevant probabilities to infer the posterior probability. Estimating the classes' prior probabilities, $P(C = c)$, from the data is relatively straightforward. However, estimating the likelihood of the features given the class, $P(F = \boldsymbol{f} \mid C = c)$, is more complicated, as the number of probabilities grows exponentially with the number of features (the curse of dimensionality). One way to sidestep the problem is to assume that features are class-conditionally independent.

Class-Conditional Independence

If class-conditional independence holds the individual features within a class are statistically independent (e.g., Domingos & Pazzani, 1997). This means that the probability of a feature configuration given a class can be factorized such that:

$$P(F = \boldsymbol{f} \mid C = c) = \prod_{j=1}^{J} P(F_j = \boldsymbol{f}_j \mid C = c) \qquad (2)$$

where $P(F = \boldsymbol{f} \mid C = c)$ denotes the likelihood of the feature configuration given the class, $P(F_j = \boldsymbol{f}_j \mid C = c)$ is the marginal likelihood of the j^{th} feature value given the class, and $j = 1, \ldots, J$ indexes the different features. Thus, according to the assumption of class-conditional independence, the likelihood of each feature value combination can be estimated from the likelihoods of the individual feature values.

Advantages The key advantage of assuming that features are class-conditionally independent is that it reduces the curse of dimensionality. For example, for 10 binary features there are 2^{10} possible feature configurations. That means, we have to estimate 1024 likelihoods of feature configurations for each class. Assuming class-conditional independence reduces the number of required likelihoods from 1024 to 8.

Another benefit is that class-conditional independence allows inferences about new feature configurations. Even if a particular combination of feature values has not been observed yet, assuming class-conditional independence allows inference of the likelihood of the feature configuration from the marginal likelihoods of the individual feature values, thereby enabling computing the posterior class probabilities.

Robustness While class-conditional independence may rarely exactly hold in real-world environments, violations of this assumption do not necessarily impair performance. For instance, a widely used classifier in machine learning is the naïve Bayes model, which treats features as class-conditionally independent and computes the posterior class probabilities accordingly. Both simulation studies and analytic results demonstrate the robustness of this model under a variety of conditions (Domingos & Pazzani, 1997). For instance, if the optimality criterion is classification accuracy (error minimization, i.e., a zero-one loss function), then even if the derived posterior probabilities do not exactly correspond to the true posterior, as long as the correct category receives the highest posterior probability, classification error will be minimized.

Summary Treating features as class-conditionally independent in a classification task can be helpful, as it simplifies the problem of parameter estimation and violations of class-conditional independence do not necessarily entail a loss in classification accuracy. On the other hand, assuming class-conditional independence also puts constraints on the types

of classification problems that can be solved. For instance, treating features as class-conditionally independent can make it impossible to solve certain classification problems, such as nonlinearly-separable category structures (Domingos & Pazzani, 1997).

From a psychological perspective, however, presuming class-conditional independence might be a plausible default assumption in category learning. If features are (approximately) class-conditionally independent, this facilitates learning and inference substantially. We designed an experiment to investigate whether people initially presume class-conditional independence, and if people change their beliefs and classification behavior when class-conditional independence does not hold in the environment.

Experiment

Our goal was to examine whether people use class-conditional independence as a default assumption in category learning when the true environmental probabilities are not known yet, that is, early in learning. In order to test this question, we designed a learning environment in which classification decisions would be strongly different if the learner presumes class-conditional feature independence, rather than basing classification decisions solely on the previous instances with the exact same configuration of feature values.

Method

Participants Thirty subjects ($M_{age} = 23, SD = 3.3$ years, 70 % females) participated in a computer-based experiment in exchange for 12 Euro.

Task Participants' task was to learn classify objects with three binary features into one out of two categories. As stimuli we used simulated biological "plankton" specimens differing in three binary features ("eye", "tail", and "claw", shown in the left image in Figure 1). The classes were labelled as "Species A" vs. "Species B". The assignment of the actual physical features and their values to the underlying probabilities, as well as the class labels, were randomized across participants.

Procedure We used a trial-by-trial supervised multiple-cue probabilistic category learning paradigm (e.g., Knowlton, Squire, & Gluck, 1994; Meder & Nelson, 2012; Nelson et al., 2010; Rehder & Hoffman, 2005). After introducing the task and familiarizing subjects with the three features, on each trial a plankton exemplar with a specific feature value combination was randomly drawn according to the true environmental probabilities (see below) and displayed on the screen. After participants made a classification decision, feedback on the true class was given and the next trial started. Learning continued until criterion performance was achieved. Criterion performance was defined as both (1) an overall classification accuracy of 98 % over the last 100 trials, and (2) accurate classification of the last five instances of every individual configuration of features.

Environment Using optimal experimental design (OED) principles (Myung & Pitt, 2009; Nelson, 2005) we conducted simulations to find environmental probabilities that best differentiate between a learner that assumes class-conditional independence and a learner that makes predictions based only on previous instances of the same feature configuration. The possible environmental probabilities for our task consisted of the following parameters: (i) the base rate of Species A (determining the Species B base rate), (ii) the likelihoods of each of the eight possible feature value combinations given Species A and (iii) the corresponding values for Species B. The parameter values were obtained via optimization, using genetic algorithms to search for desirable environments which had frequent configurations of features with large absolute discrepancies between the actual posterior probability of Species A, and the posterior probability presumed based on the class-conditional independence assumption. Formally, the genetic algorithm optimized the following fitness function:

$$\sum_{i=1}^{I} [P_{\text{true}}(C = c \mid F = \boldsymbol{f}_i) - P_{\text{cci}}(C = c \mid F = \boldsymbol{f}_i)]^2 \times P(F = \boldsymbol{f}_i)^2 \tag{3}$$

where i indexes all possible feature value combinations and the subscripts true vs. cci indicate the posteriors calculated according to the true vs. class-conditionally independent parameters.

The obtained environment is summarized in Figure 1. The environment contains five out of eight possible feature combinations (henceforth denoted as 111, 000, 100, 010, 001); the remaining three combinations (011, 101, 110) do not occur. The figure illustrates the category base rates, the likelihoods of the feature configurations given the two classes, as well as the marginal likelihoods of the features, which provide the basis for inferring posterior probabilities according to the class-conditional independence assumption. Note that although nothing in the optimization prescribed finding a deterministic environment, in fact the posterior probabilities of Category A are one or zero, for each of the feature configurations that occurs.

In this environment, assuming class-conditional independence leads to classification decisions that systematically deviate from decisions based on the true environmental probabilities. Table 1 summarizes the feature configurations, their probability of occurrence, the posterior probabilities according to the true environmental probabilities, and the posterior probabilities derived assuming class-conditional independence. For four out of the five feature configurations, the classification decision derived assuming class-conditional independence conflicts with the actual class membership (indicated by $\neq$ in Table 1).

Consider feature configuration 111. This item always belongs to Species A in the true environment. If features are treated as class-conditionally independent, it belongs to Species A with probability 0.91. The small difference between the actual probability of 1.00 and 0.91 should not

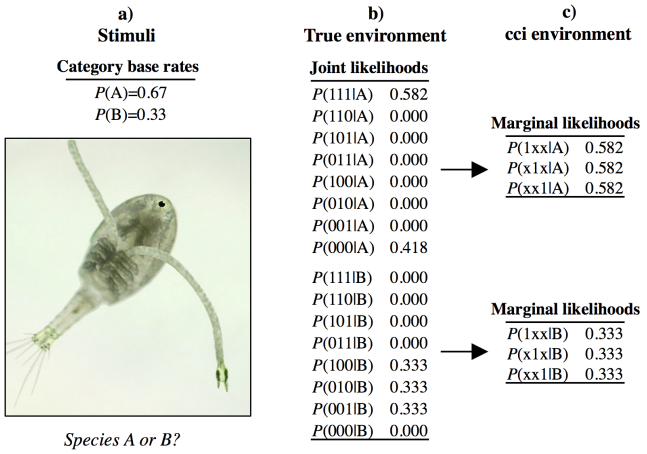

Figure 1: Task environment. a) Stimuli and base rates of classes. b) Joint likelihoods of true environment. c) Marginal likelihoods used assuming class-conditional independence

change the learner's classification decision for this stimulus. This, however, is not true for the other items. For instance, according to the true environment, item 000 belongs to Species A with probability 1, but assuming class-conditional independence entails that it belongs to Species B with probability 0.67. Thus, a learner assuming class-conditional independence would believe that on average about 67 % of the 000 items belong to Species B, despite experiencing that it always belongs to Species A. The same divergence holds for the other three configurations (100, 010, 001): whereas all of those items actually belong to category B, treating features as class-conditionally independent entails that the probability for category A is higher (0.58).

Table 1: True environment vs. assuming class-conditional independence (cci).

Features	P(features) true env	P(class | features) true env		with cci
1 1 1	0.39	A 1	=	A 0.91
1 0 0	0.11	B 1	$\neq$	A 0.58
0 1 0	0.11	B 1	$\neq$	A 0.58
0 0 1	0.11	B 1	$\neq$	A 0.58
0 0 0	0.28	A 1	$\neq$	B 0.67

The strongest discrepancy is for the 000 configuration, which is the second-most-frequent configuration, occurring with probability .28. Note that a hypothetical learner (even with perfect memory) who assumes class-conditional independence of features, and is unable to give up this assumption, will never learn the true statistical structure of this environment, even after completing a quadrillion learning trials.

Achieving criterion performance would also be impossible if learners looked at one feature only (at 1xx, or x1x, or xx1 and ignoring the x). Considering single features, participants should think any feature configuration belongs to Species A with probability 0.78. This holds for attending solely to any of the three features.

Hypotheses

If participants make no (not even tacit) assumptions of class-conditional feature independence, and learn each item separately, then items could be learned in order of their frequency of occurrence (a *frequency-of-configuration hypothesis*). If participants approach the task by assuming features to be class-conditionally independent, classification decisions should systematically deviate from ones derived from the true environmental probabilities, especially early in learning (a *posterior-discrepancy hypothesis*).

Both hypotheses predict the fewest errors for item 111, the most frequent feature configuration and the one for which the class-conditional independence posterior is closest to accurate. For the four critical items, the difference in posterior probability is the largest for item 000. The posterior-discrepancy hypothesis predicts the most errors for item 000, and thus that the ordering of errors should be 111<100≈010≈001<000. However, the frequency-of-configuration hypothesis predicts that the ordering of classification errors should be 111<000<100≈010≈001.

Key empirical questions are therefore whether there are any systematic differences in learning rate for the individual items, whether the early learning data suggest a presumption of class-conditional independence, and if so, whether the occurrence frequency of an item or the degree to which class-conditional independence fails on it determine learning.

Results and Discussion

All participants reached criterion performance, i.e. learned the category structure (in a mean number of 391 trials, SD=155, Md=348, range 210 to 808 trials). To reach criterion performance, participants needed to classify each individual feature configuration correctly five times in a row. To investigate whether there was a difference in learning speed for the different feature configurations, we calculated the number of times each item needed to be observed before reaching this criterion (Table 2). We will first consider learning time and then error rates.

Table 2: Number of trials an item needed to be seen to correctly classify it five times in a row.

Features	Trials mean	SD (SE)	median
1 1 1	10.4	10.7 (1.9)	7.0
1 0 0	11.4	8.0 (2.1)	7.5
0 1 0	11.5	7.7 (2.1)	9.0
0 0 1	11.5	7.0 (2.1)	9.0
0 0 0	15.8	11.5 (2.9)	13.5

In our data most subjects learned item 111 before item 000 (22 out of 30, binomial $p < .02$), which is consistent with both hypotheses. Did learning time follow

the frequency-of-configuration hypothesis, or the posterior-discrepancy hypothesis? The posterior-discrepancy hypothesis predicts an ordering of 111<100≈010≈001<000, whereas the item-frequency hypothesis's ordering prediction is 111<000<100≈010≈001. The critical difference in predictions is between the learning time for items 100, 010, and 001 and item 000. The frequency hypothesis predicts that item 000 will be learned faster, whereas the posterior discrepancy hypothesis predicts that items 100, 010, and 001 will be learned first. Here, our results strongly support the posterior discrepancy hypothesis, and contradict the item frequency hypothesis. Items 100, 010 and 001 were learned more quickly by more people than item 000, despite item 000's greater frequency (item 001 faster: 21 out of 30, binomial $p < .05$; item 010 faster: 20 out of 30, binomial $p < .1$; item 100 faster: 21 out of 30, binomial $p < .05$). Moreover, there was a non-significant trend for items 100, 010, and 001 to take longer than item 111; consistent with the posterior discrepancy hypothesis but not the configuration frequency hypothesis.

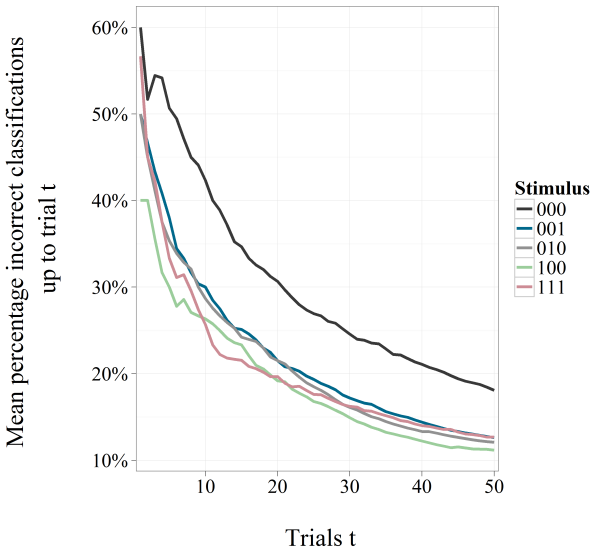

Figure 2: Percentage incorrect classifications for the first 50 trials each item was encountered.

The error rates throughout early learning are summarized in Figure 2. This figure corroborates the analysis of the number of learning trials required for each stimulus configuration: item 000 was clearly the most difficult to learn. As this feature configuration is the one for which the difference in posterior probability is largest when assuming class-conditional independence versus using the full true environmental probabilities, this finding is consistent with the idea that people treat features as being class-conditionally independent early in learning. However, items 100, 010 and 001 were much closer to (or even indistinguishable from) item 111, consistent with the above analysis in Table 2.

General Discussion

The present paper examined the role of the assumption of class-conditional independence of features in category learning. While different types of conditional independence assumptions play an important role in various scientific debates and computational models of cognition, little is known about their descriptive validity in the context of classification learning with multiple cues. Our goal was to empirically investigate whether people initially (early in learning) treat features as class-conditionally independent. The present results partially support the idea that people initially treat features as class-conditionally independent and make classification decisions accordingly. We think of the results as tentative because some aspects of the data are not perfectly clear.

Our focus in the present study was on participants' behavior early in learning, when evidence about the category structure and environmental probabilities is limited. This approach is similar to the studies of Smith and Minda (1998), who investigated possible transitions in categorization strategies and stimulus encoding over the course of learning.Their finding was that late in learning exemplar models (e.g., Medin & Schaffer, 1978) accounted best for subjects' behavior, but that this was not the case early in learning (in which a prototype model seemed to better account for human performance, see below). This is also a possible explanation for the finding that despite strongly violating class-conditional independence, the environment in our experiment was clearly learnable. Participants could have initially treated features as class-conditionally independent and computed posteriors accordingly and later shifted to an exemplar-based strategy to minimize classification error.

A key methodological aspect of our study was to use optimal experimental design principles to find environments that would allows us to directly test whether people use class-conditional independence as a default assumption in categorization. Interestingly, the optimizations told us that the best environment to differentiate between a learner that assumes class-conditional independence and a learner that makes predictions based only on previous instances of the same feature configuration was deterministic. The crucial aspect of this environment, however, is not that it is deterministic, but that it entails a nonlinearly separable category structure. Since the class-conditional independence model induces a linear decision bound (Domingos & Pazzani, 1997), it could not achieve criterion performance in this particular task environment.

This, in turn, relates our study to earlier research in psychology, which investigated whether linearly separable categories are easier to learn than nonlinearly separable ones (e.g., Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981). This research focused on two types of categorization models, exemplar- and prototype-models, both of which assume that categorization decisions are derived from similarity comparisons (either to specific exemplars stored in memory or to prototypes of categories). By contrast, we investigated category learning and human subjects' initial assump-

tions from the perspective of probabilistic inference (see also Anderson, 1991; Fried & Holyoak, 1984), a conceptually different view. Nevertheless, there are some interesting connections between our work and these earlier (similarity-based) models. For instance, assuming class-conditional independence entails that not all information (about feature configurations and corresponding class probability) is encoded during learning, but only marginalized conditional likelihoods and category base rates. In this respect the class-conditional independence model is similar to prototype models, which encode parametric information of central tendencies (e.g., mean or mode of feature values) that form the prototype (e.g., Smith & Minda, 1998).

Importantly, these accounts assume that information is stored separately for each feature and the to-be-classified item is compared to the prototypes separately on each feature dimension individually. Conversely, a learner who makes no assumptions about the structure of the relations between classes and features and directly tracks the true environmental probabilities is conceptually more similar to exemplar models of category learning. The difference is that prototype models, like our independence model, do not need to store each individual instance that is experienced.

In sum, the current paper adds to the debate about the role of conditional independence assumptions for computational models of cognition. The task environment identified based on optimal experimental design principles allowed us to directly examine the descriptive validity of this assumption in category learning. Here, we do find evidence consistent with its use.

Acknowledgments

This research was supported by grants NE 1713/1 to JDN and ME 3717/2 to BM, from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). We would like to thank Gregor Caregnato for data collection and Laura Martignon, Michael Waldmann, Ralf Mayrhofer, and the reviewers for their helpful comments. We also thank Jorge Rey and Sheila O'Connell (University of Florida, FMELe) for allowing us to base our artificial plankton stimuli on their copepod photographs.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *3*, 409–429.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Fried, L. S. & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *2*, 234–257.

Hertwig, R., Barron, G., Weber, E., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Knowlton, B., Squire, L., & Gluck, M. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*, 106–120.

Mayrhofer, R., Hagmayer, Y., & Waldmann, M. (2010). Agents and causes: A Bayesian error attribution model of causal reasoning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Meder, B. & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*, 119–148.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *3*, 207–238.

Medin, D. L. & Schwanenflugel, P. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, 355–368.

Movellan, J. R. & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113–148.

Myung, J. & Pitt, M. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 832–840.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999.

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters. *Psychological Science*, *21*, 960–969.

Pearl, J. (2000). *Causality. Models, Reasoning and Inference*. New York: Cambridge University Press.

Rehder, B. & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264–314.

Rehder, B. & Hoffman, A. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

Smith, J. D. & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. New York: Springer.

Von Sydow, M., Meder, B., & Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Walsh, C. R. & Sloman, S. A. (2007). Updating beliefs with causal models: Violations of screening off. In J. R. A. M. A. Gluck & S. M. Kosslyn (Eds.), *Memory and Mind: A Festschrift for Gordon H. Bower*.