Kruschke, J. K. & Denton, S. E. (2010). Backward blocking of relevance-indicating cues: evidence for locally Bayesian learning. In: C. J. Mitchell & M. E. Le Pelley (Eds.), *Attention and Associative Learning: From Brain to Behaviour*, pp. 273-304. Oxford, UK: Oxford University Press. ISBN 9780199550531

Chapter 11

Backward blocking of relevance-indicating cues: Evidence for locally Bayesian learning

John K. Kruschke and Stephen E. Denton

The phenomenon in associative learning called 'blocking' has been a central target for theories that aim to explain learning and attention. The training phases of the blocking procedure, described in detail below, can be run in backward order, yet still produce the blocking effect (e.g., Shanks,1985; Dickinson and Burke,1996; Kruschke and Blair, 2000). Explaining forward and backward blocking, along with other associative learning phenomena, has proven to be challenging. Some accounts of forward blocking include an attentional mechanism that modulates the influence of the cues on learning and/or responding (e.g., Kruschke, 2001; Mackintosh, 1975). Some accounts of backward blocking employ a Bayesian framework in which different combinations of associative weights are considered simultaneously, with more belief allocated to the combination that is most consistent with training items (e.g., Dayan and Kakade, 2001; Tenenbaum and Griffths, 2003).

A theoretical framework that is able to combine the attentional and Bayesian approaches is called 'locally Bayesian learning' (Kruschke, 2006b). The framework is based on the idea that a learning system may consist of a sequence of subsystems in a feed-forward chain, each of which is a locally Bayesian learner. The argument for locally-learning layers was as follows. First, Bayesian learning is very attractive for explaining retrospective revaluation effects such as backward blocking, among many other phenomena (Chater, Tenenbaum, and Yuille, 2006). Second, globally Bayesian learning may also be unattractive for a number of reasons. In a large learning system, there are too many combinations of parameters to keep track of in a monolithic joint parameter space. Furthermore, many globally Bayesian models do not explain learning phenomena (such as 'highlighting', Kruschke, 2010) that depend on training order, because the models treat all training items as equally representative of

the world to be learned, regardless of when the items occurred. Finally, the level of analysis for theories of learning is arbitrary: Learning occurs simultaneously at the levels of neurons, brain regions, functional components, individuals, committees, institutions, and societies, all of which may be modeled (in principle, if not accurately) as Bayesian learners. Therefore, a system of locally Bayesian learning components may retain some attractions of Bayesian models while also implementing Bayesian learning in smaller, tractable parameter spaces.

The general framework for locally Bayesian learning has been instantiated in a particular two-layer model, wherein one layer learns how to allocate attention to cues, and a second layer learns how to associate attended cues with outcomes (Kruschke, 2006b). The model showed retrospective revaluation effects such as backward blocking while also showing the order-sensitive phenomenon of highlighting. The two-layer model had locally Bayesian learning in both of its layers. Bayesian learning was needed in the upper, outcome layer to produce effects such as backward blocking. Attentional shifting was needed to produce effects such as highlighting. But there was no phenomenon that demanded Bayesian learning in the lower, attentional layer. Bayesian learning was conducted in the lower layer merely for mechanistic consistency and as a demonstration of the more general framework for locally Bayesian learning.

The purpose of the present chapter is to report a novel learning design, the results of which suggest the need for Bayesian learning in the lower, attentional layer. In essence, the results show backward blocking of cues to attentional allocation, as distinct from traditional designs that show backward blocking of cues to response allocation. This pattern of results can be accommodated by a model that has layers of locally Bayesian learning.

Background

Forward and backward blocking

In the standard forward blocking procedure, a learner is initially trained with cases of cue A leading to outcome X, denoted A→X. Subsequently, training continues with cases in which two cues, A and B, lead to the same outcome X, denoted A·B→X. It turns out that the strength of association from cue B to outcome X is weaker than if the initial training with A had not occurred. In other words, the learning of the association from B to X has been 'blocked', or attenuated, by the previous learning of the association from A to X. The phenomenon of blocking, first reported by Kamin (1968), challenges theories that base strength of association on merely the number of co-occurrences of cue and outcome (but cf. Miller and Matzel, 1988).

There are many types of explanations of blocking, but one family of explanations posits a role for attention. The intuition is that when learning cases of A · B→X, the learner re-allocates attention away from the redundant cue B because it is distracting resources away from the cue A that is already known to generate the correct outcome. In other words, the learner learns to ignore cue B. Some evidence for the attentional explanation comes from studies of learning about cue B after blocking. If learners have learned to ignore B, then subsequent learning about it should be retarded. This prediction has been confirmed (e.g., Kruschke and Blair, 2000; Kruschke, 2005; LePelley, Beesley, and Suret, 2007; Mackintosh and Turner, 1971; Mitchell, Harris, Westbrook, and Griffiths, 2008). Other evidence for reduced attention to the blocked cue comes from eye tracking experiments, in which it has been shown that gaze duration is reduced for blocked cues (Kruschke, Kappenman, and Hetrick, 2005; Wills, Lavric, Croft, and Hodgson, 2007).

Backward blocking, as a training procedure, simply reverses the phases of training in standard forward blocking. In other words, learners are first trained with cases of A · B→X, and subsequently trained with cases of A→X. It turns out that the strength of association from B to X is again weakened by the A→X training, even though it happened after the training with B, and even though B never appeared in the subsequent training (Shanks, 1985). Thus, cue B seems to have been retrospectively revalued in its absence. Backward blocking and other retrospective revaluations are especially challenging to theories of associative learning (for reviews see DeHouwer and Beckers, 2002; Dickinson, 2001).

The attentional theories that account for blocking do not account for backward blocking. This failure is caused by the fact that the theories rely on the presence of a cue to learn how much to attend to it. If a cue is absent, the models do not change its attention strength.

One class of theories that accommodates backward blocking is Bayesian models of association (e.g., Dayan and Kakade, 2001; Sobel, Tenenbaum, and Gopnik, 2004; Tenenbaum and Griffiths, 2003).[1] These Bayesian theories posit a set of associative hypotheses simultaneously entertained by the learner. For illustration, suppose that the learner considers three hypotheses: (1) A indicates X and B is irrelevant, (2) B indicates X and A is irrelevant, and

[1] A class of non-Bayesian theories of backward blocking asserts that absent-but-expected cues acquire reduced associations when the expected outcome is present. For absent-but-expected cues, either the learning rate or encoding of the absent-cue is negative, resulting in a loss of association (Dickinson and Burke,1996; Markman, 1989; Tassoni, 1995; Van Hamme and Wasserman, 1994).

(3) either A or B indicate X.The three hypotheses are mutually exclusive, and exhaust the space of possibilities for this particular learner. After seeing the initial cases of $A \cdot B \rightarrow X$, all three hypotheses have some credibility. But after seeing cases of $A \rightarrow X$, the first and third hypotheses gain credibility, because they are both consistent with the additional training. Because the set of hypotheses are mutually exclusive and exhaustive, when the first and third hypotheses gain credibility, the second hypothesis loses credibility. Therefore, across all the hypotheses, there is reduced strength of belief that B indicates X. For a detailed tutorial, see the discussion of Bayesian associative models by Kruschke (2008).

Bayesian approaches to learning are attractive for a variety of other reasons. Bayesian models are not limited to associative formalisms, but can instead incorporate complex structural representations into the hypothesis space. Bayesian models merely assume that the learner executes normatively correct learning (i.e., uses Bayes' rule) on whatever formal hypothesis space is posited. This representational flexibility allows Bayesian models to be applied to situations from learning by neurons (Deneve, 2008) to learning of language (Xu and Tenenbaum, 2007).

Locally Bayesian learning

Bayesian formalisms are very attractive as theories of learning, but implementing them can be difficult because of their computational complexity. In principle, Bayesian systems need to keep track of the credibility of every possible hypothesis. In hypothesis spaces with many parameters, such as numerous associative weights, the system needs to keep track of the credibility of every possible combination of parameter values in a high-dimensionality joint parameter space. In some simple models, this can be done exactly and easily because the entire distribution across beliefs can be summarized by a simple function such as a multivariate normal distribution. In more complicated models, the infinite space of hypotheses can be represented by a large but finite random sample of representative hypotheses. In these 'particle filter' approximations, as new cue outcome cases are experienced, the representative hypotheses are resampled to reflect the new experience. These methods are not trivial, however, and Bayesian computation can be quite challenging in large hypothesis spaces.

One way to attack the problem is to recognize that learning happens at many levels of analysis simultaneously. Neurons learn, brain regions learn, functional components of mind learn, individual people learn, teams of people learn, and so on. Indeed, even more microscopic and macroscopic systems may learn. Any of those levels of analysis may be amenable to description as Bayesian learning.

Therefore, it is at least plausible that component processes of the mind may be describable as Bayesian learners, because the components have a tractable hypothesis space. Whether or not the system as a whole is Bayesian depends on how the components interact. A scheme for interaction of hierarchically organized, locally Bayesian learners was described by Kruschke (2006b).

The general framework for locally Bayesian learning assumes that there are component processes in a hierarchy from stimulus encoding to response generation. Each functional component in the hierarchical chain takes its local input representation, transforms it, and delivers its local output representation to the next component in the chain. The transformation is parameterized, meaning that the exact quantitative behaviour of the local transformation depends on its parameter values. As a simple example, consider a linear transformation $y = mx + b$, where x is the input and y is the output. The slope m and the intercept b are parameters that govern the quantitative value of the output for any given input. The parameters are learned as exemplary $\langle x, y \rangle$ values are experienced. The learning of parameters is Bayesian reallocation of credibility to combinations of parameter values that are most consistent with the incoming stimuli and the target response. For example, if the component system has experienced $\langle x, y \rangle$ pairs such as $\langle 1, 2.01 \rangle$, $\langle 2, 3.99 \rangle$, and $\langle 3, 6.01 \rangle$, then it will allocate strong credibility to $m = 2$ and $b = 0$, i.e., $y = 2x + 0$, and weaker credibility to other parameter values.

The challenge for such a framework is determining the target response for interior components, because the outside world only indicates the target response for the final component that generates an overt response. Formally, the world only supplies the exterior stimulus x and the exterior target response y. For a component buried in the interior of the hierarchy, the component's input can be computed by propagating the exterior stimulus x up through the transformations leading into the component. The component's target output, however, is not obvious, because the transformations feed forward from x to y, not the opposite direction.

A heuristic for determining interior targets is as follows. The target for a component should be whatever is the input to the next component that maximizes the probability of achieving the target of that next component. Start at the final component and determine the input to that component that would maximize the probability of achieving the known exterior target. Use that input as the target for the penultimate component. Given that target for the penultimate component, determine the input to the penultimate component that would maximize the probability of achieving its target. Use that input as the target for the preceding component. Continue this process down the hierarchy of components until every component has a target.

Notice that the targets selected in this manner are the inputs that are most consistent with the component's current beliefs. In other words, at any given time, the component gives strongest credibility to some particular combinations of parameter values. That component finds the candidate input that would be most consistent with the parameter values that the component currently believes. That best candidate input is declared to be the target for the preceding layer. In this way, each component is teaching the preceding component to 'tell me what I want to hear' based on its current beliefs.

A crucial consequence of choosing interior targets this way is that the temporal order of data has an effect on what is learned internally. The reason is that the interior targets are determined by the components' current beliefs, which depend on the set of data experienced so far. The interior targets generated for any particular exterior input-output pair depend on what has been previously learned.

When a component transformation has a target, along with its input, then the parameters of the component transformation are adjusted via Bayesian learning. Parameter values that are consistent with the input and target are deemed more credible. There are various temporal dynamics that could be used for interleaving parameter learning and target determination (Kruschke, 2006b, p. 683, footnote 2).

Attention in locally Bayesian learning

Locally Bayesian learning is a general framework for models. One particular application is to attention in associative learning. In this application, the first component learns how to allocate attention across cues. Attention is an interior value, and is not explicitly provided by exterior data. The second component learns how to associate attentionally-filtered cues to outcomes. Both the attentional allocations and the outcome associations are acquired via locally Bayesian learning.

Previous work demonstrated the usefulness of this architecture (Kruschke, 2006a, 2006b). In particular, locally Bayesian learning of the output associations allows the system to accommodate backward blocking and other retrospective revaluation phenomena. Figure 11.1 shows the basic elements of locally Bayesian learning applied to attentional learning. The lowest nodes encode the cues, the middle nodes represent the attentionally filtered cues, and the upper nodes represent the (anticipated) outcome. The left side of Figure 11.1 illustrates what happens when the system is trained in the backward blocking procedure. In this procedure, the model first experiences cases of $A \cdot B \rightarrow X$, and then experiences cases of $A \rightarrow X$. The upper-left part of the figure shows a variety of output–weight combinations that the model finds

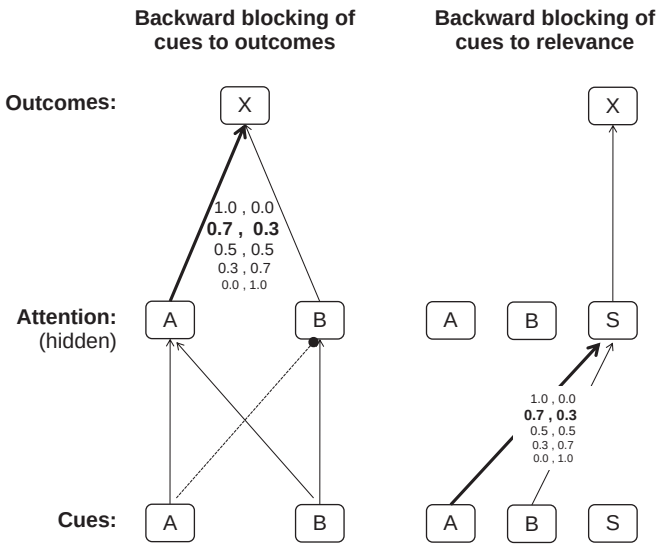

Fig. 11.1 Schematic of locally Bayesian learning applied to backward blocking. *Left*: Backward blocking of cues to outcomes. *Right*: Backward blocking of context cues to relevance. The pairs of numbers suggest some of the weight combinations that the model finds credible.

credible after training. The size of the font suggests how strongly the model believes in the weight combination. Thus, the model gives greatest credibility to an associative weight of 0.7 from A and an associative weight of 0.3 from B. But the model also gives some modest credibility to other weight combinations that are reasonably consistent with the training items. The layer of weights leading into the attentional gates also has a distribution of credibility across possible weight combinations.

Attentional shifting, and the temporal dependency of the scheme for selecting interior targets, also allows the system to accommodate the highlighting phenomenon (Kruschke, 2010). Highlighting is a trial-order effect that is particularly vexing for Bayesian models that treat all trials as equally representative of the world. The temporal dynamics of locally Bayesian learning also let the system show other trial-order effects, such as stronger forward blocking than backward blocking. A variety of other phenomena are addressed by the model.

The ability of the model to show backward blocking (and other effects) relied on Bayesian learning in the upper associative layer. But none of the phenomena considered by Kruschke (2006b) demanded Bayesian learning in the lower, attentional layer. In principle, the attentional layer could

have been a non-Bayesian learner, and all the same effects could have been produced.

The primary goal of the present article is to report a phenomenon that does suggest the need for Bayesian learning of attentional allocation. The argument goes like this: Backward blocking is naturally accounted for by Bayesian learning. If it can be shown that there is backward blocking in the learning of attentional allocation, then one candidate explanation is that there is Bayesian learning of attentional allocation.

The right side of Figure 11.1 illustrates what is meant by backward blocking of attentional allocation. The complete design of the cue–outcome combinations will be described later, but the gist is provided here. Contrary to the standard design, cues A and B are not themselves diagnostic of the outcome: Across trials, the outcome occurs just as often when cues A and B are absent as when they are present. A different cue, labeled S in Figure 11.1, is perfectly predictive of the outcome, but only when cue A or cue B is present. When cues A and B are absent, cue S is not predictive. Thus, cues A and B do not indicate what outcome to anticipate, but they do indicate what other cue is relevant. In essence, cues A and B indicate what other cues should be attended.

The goal of the experiment reported below is demonstrate backward blocking of such cues to relevance. In the first part of training, cues A and B are both present whenever S is diagnostic. Therefore the model should learn to attend to S in the presence of A and B. Later in training, cue A is present without cue B whenever S is diagnostic. If there is backward blocking of B, then the association from B to S should be weakened. The weight pairs in the lower-right side of Figure 11.1 are intended to suggest the credible weight combinations after backward blocking of B as a cue to relevance.

The purpose of the experiment and modelling presented here is not to rule out other explanations or disconfirm other models. Instead, the goal is to bolster an assumption of locally Bayesian learning applied to attentional learning, a model which was already shown to address a spectrum of phenomena (Kruschke, 2006b). There are surely other models that can accommodate the data from the one new experiment presented here, but for a model to compete with locally Bayesian learning, the candidate model should also accommodate the spectrum of other phenomena addressed by locally Bayesian learning.

The new experiment presented here is offered merely as a suggestive proof of concept. Future experiments will be necessary to generalize the conditions under which the effects are observed, and to rule out alternative explanations. Nevertheless, it is hoped that the experiment and modeling may provoke interesting new ideas and research. In particular, the experiment presented here

might not have been invented were it not for the implications of locally Bayesian learning applied to attentional learning.

Experiment: Blocking and backward blocking of cues to relevance

Different cues can be relevant in different contexts. For example, when driving an automobile, the colour of the stoplight is relevant to the decision to stop or go, but when walking, the colour of the pedestrian signal is relevant to the decision to stop or go. In general, the cues in an environment can suggest which sources of information are relevant for determining a response. For example, the cue of having a steering wheel in your hands does not tell you whether to stop or go, but it does indicate that you should attend to stoplights, which will indicate whether you should stop or go.

Cues that are indirectly informative, such as the steering wheel in the previous example, are a type of context cue. The term 'context' has no generally accepted technical definition. Sometimes context refers to stimulus attributes that are spatially ambient (not focal) or temporally extended (i.e. tonic not phasic). Other times, context refers to cues that can be focal and phasic but that are not directly correlated with outcomes. This latter character is emphasized here. There has not been a vast amount of previous research into the role of context in learned attentional allocation, but several lines of work have indicated that people can learn about 'irrelevant' context as a cue to Attention (e.g., Chun, 2000; Nelson, 2002; Rosas, Callejas-Aguilera, Ramos-Alvarez, and Abad, 2006; Yang and Lewandowsky, 2003).

The present experiment is aimed at demonstrating forward and backward blocking of contextual cues to relevance. Continuing the example from driving discussed earlier, the idea is that people first experience steering wheels as a cue to attend to stoplights, and later people experience steering wheels along with a newly installed car stereo as a compound cue to attend to stoplights. The association from car stereo to attention may be blocked because of the previously learned association from steering wheel to attention.

To test this idea, we conducted a series of experiments in which all the cues were simple words on a computer screen, such as 'radio' and 'ocean'. Some words were perfectly correlated with the correct key to press. Other words had zero correlation with the correct response key, but were perfect indicators of which other words on the screen were relevant to the choice of response key. For example, suppose that people have learned that radio indicates key X and ocean indicates key Y. Subsequently, both 'radio' and 'ocean' appear simultaneously. Should the response be X or Y? The conflict is resolved by a third

word on the screen, e.g., 'queen', which indicates to attend to 'radio'. Our experiments revealed that it was difficult for people to learn this sort of contextual dependency in a brief (< 20 min) experiment when the cues had no structural or semantic indicators of which were context cues and which were response cues. For the few people who could learn such structures quickly, we observed signs of blocking and backward blocking of cues to relevance. But it was unsatisfying to base conclusions on a small subset of participants. Presumably, all people could learn such structures if given enough practice, but before subjecting people to endurance training, we explored other stimulus arrangements.

In order to facilitate learning, and for purposes of an experiment that can act as a proof of concept, we set up a cue arrangement in which it was natural to think of some cues being indicators of responses, and other cues being indicators of which response cues to attend to. The learners were instructed that they were to diagnose the fictitious disease associated with symptoms, but only indirectly, by learning which medical specialist knew about which symptoms. For example, a patient might have the symptoms heartburn and myalgia, for which Specialist 1 says the patient has disease F, but Specialist 2 says the patient has disease J. After the learner makes a guess about the disease, corrective feedback indicates that it was disease J, thereby implying that Specialist 2 knows about these symptoms. The learner should therefore learn that when symptoms heartburn and myalgia occur, s/he should attend to Specialist 2, and give the response stated by Specialist 2. Importantly, there is no correlation between symptoms and diseases across trials. For example, half the time that heartburn and myalgia occur, the correct disease is F, but half the time the correct disease is J. Specialist 2 always indicates the correct diagnosis, however, for these symptoms.

The cues that indicate the correct overt response are here called 'response cues'. In the present scenario, the medical specialists are the response cues. Other cues that indicate which response cues to attend to are here called 'context cues'. In the present scenario, the symptoms are the context cues. These appellations, i.e., response cue versus context cue, are potentially misleading. On the one hand, the so-called context cues do indicate a response, but that response is an essentially covert re-allocation of attention (which may or may not have overt signatures such as eye movements or other orienting responses). On the other hand, the so-called response cues need not be known in advance to be indicators of overt responses; the response cues might serve as context to other cues. Despite these infelicities of nomenclature, a key aspect of intuitively contextual information *is* captured by the 'context' cues: They are not

directly predictive of the correct overt response. The context cues only indicate which response cues to attend to, and the response cues, in turn, indicate which overt response to make.

Method

Design

Table 11.1 shows the design components of the experiment. Each phase has a different arrangement of context cues. In the Single Context phase, a single context cue is present with two Specialists who give conflicting diagnoses. For example, a trial might consist of $A_{S1}S1_FS2_J \rightarrow F$, which means that context cue A_{S1} occurred with specialists $S1_F$ and $S2_J$, with correct diagnosis F. The subscripts on the cues denote what the cue is intended to indicate. The notation A_{S1}, for example, means that cue A indicates specialist S1. This correspondence had to learned, however. Notice that across the eight cases of the Single Context phase, the context cue A occurs twice with outcome F and twice with outcome J. Hence the context cue is uncorrelated with the correct diagnosis. In all cases of the Single Context phase, only a single context cue occurs.

In the Redundant Context phase, some cases had two context cues. In particular, context cues A and B occurred together, and context cues C and D occurred together. As these context cues never occurred separately, they

Table 11.1 Components of the experiment design

Phase	Items	
Single Context	A_{S1} $S1_F$ $S2_J \rightarrow F$	A_{S1} $S1_F$ $S3_J \rightarrow F$
	A_{S1} $S1_J$ $S2_F \rightarrow J$	A_{S1} $S1_J$ $S3_F \rightarrow J$
	E_{S3} $S3_F$ $S2_J \rightarrow F$	E_{S3} $S3_F$ $S1_J \rightarrow F$
	E_{S3} $S3_J$ $S2_F \rightarrow J$	E_{S3} $S3_J$ $S1_F \rightarrow J$
Redundant Context	A_{S1} B_{S1} $S1_F$ $S2_J \rightarrow F$	$A_{S1}B_{S1}$ $S1_F$ $S3_J \rightarrow F$
	A_{S1} B_{S1} $S1_J$ $S2_F \rightarrow J$	A_{S1} B_{S1} $S1_J$ $S3_F \rightarrow J$
	$C_{S2}D_{S2}$ $S2_F$ $S1_J \rightarrow F$	$C_{S2}D_{S2}$ $S2_F$ $S3_J \rightarrow F$
	C_{S2} D_{S2} $S2_J$ $S1_F \rightarrow J$	$C_{S2}D_{S2}$ $S2_J$ $S3_F \rightarrow J$
	E_{S3} $S3_F$ $S2_J \rightarrow F$	E_{S3} $S3_F$ $S1_J \rightarrow F$
	E_{S3} $S3_J$ $S2_F \rightarrow J$	E_{S3} $S3_J$ $S1_F \rightarrow J$
Test: Conflicting Context	A_{S1} C_{S2} $S1_F$ $S2_J \rightarrow ?$	A_{S1} C_{S2} $S1_J$ $S2_F \rightarrow ?$
	A_{S1} D_{S2} $S1_F$ $S2_J \rightarrow ?$	A_{S1} D_{S2} $S1_J$ $S2_F \rightarrow ?$
	B_{S1} C_{S2} $S1_F$ $S2_J \rightarrow ?$	B_{S1} C_{S2} $S1_J$ $S2_F \rightarrow ?$
	B_{S1} D_{S2} $S1_F$ $S2_J \rightarrow ?$	B_{S1} D_{S2} $S1_J$ $S2_F \rightarrow ?$

Note: An item is shown in the format, Cues→Correct Response. The subscripts on the cues Indicate the design's intended correspondence from that cue. Cues A_{S1}, B_{S1}, C_{S2}, D_{S2}, and E_{S3} are symptoms. Cues $S1_F$, $S1_J$, $S2_F$, $S2_J$, $S3_F$, and $S3_J$ are specialists. Responses F and J are disease labels.

are called 'redundant'. As in the single-context phase, all context cues are uncorrelated with the correct diagnosis.

Notice that when the single-context phase occurs before the redundant-context phase, the context cues instantiate a standard blocking sequence. People learn first that context cue A indicates specialist 1, and then people see that context cue A with context cue B also indicate specialist 1. There may be blocking of learning about the redundant cue B. When the redundant-context phase occurs before the single-context phase, this constitutes a backward blocking design for the context cues.

Blocking is assessed in the test phase, wherein conflicting cues appear together. In particular, cue A_{S1} appears with either C_{S2} or D_{S2}, and cue B_{S1} appears with either C_{S2} or D_{S2}. If context cue B_{S1} is blocked, then when it is paired with either C_{S2} or D_{S2}, the response appropriate to Specialist 2 should be given, in preference over the response appropriate to Specialist 1. As a partial test that Specialist 2 is not favored generically whenever conflict occurs, the cases in which A_{S1} appears with either C_{S2} or D_{S2} should favor Specialist 1, or at least not Specialist 2. This issue is addressed further in the discussion after the results are reported.

There were three different 'epochs' of the experiment. Each epoch had the phases shown in Table 11.1, in different orderings or with different types of feedback. The first epoch was a backward blocking sequence (redundant-context phase before single-context phase), but with the correct specialist explicitly and directly indicated by corrective feedback. This design constituted a replication of the backward blocking design of Kruschke and Blair (2000, Experiment 3), except that the present procedure used a shorter-duration training to criterion(details of which are described below). The hope was that by replicating a previous design known to produce backward blocking, we would observe the effect here too.

The second epoch was a forward blocking sequence(single-context phase before redundant-context phase), with the correct specialist indicated only indirectly and implicitly via the correct diagnosis. In other words, this epoch was trained as shown in Table 11.1. New symptoms were used in all epochs, so that novel learning was involved. As forward blocking tends to be more robust than backward blocking (e.g., Kruschke and Blair, 2000; Shanks, 1985), it was hoped that we would be able to observe forward blocking even in this complex, indirect-feedback situation. This sort of procedure has never been reported before, as far as we know.

The third epoch was a backward blocking sequence (redundant context before single context), with the correct specialist indicated indirectly via the correct diagnosis. In other words, it was just like the second epoch, but

with the phases of training reversed. This epoch constituted the main focus of the experiment. We hoped to observe backward blocking of context cues. Notice that the stimulus–outcome structure of Epoch 3 is identical to that of Epoch 1; only the response and feedback are different between the two epochs.

Procedure

Table 11.1 shows one 'block' of trials in each phase. Within each block, the cases were presented in a random order. Training continued in each phase until accuracy in a block exceeded 87% correct, meaning at least 7 correct in 8-trial blocks or at least 11 correct in 12-trial blocks. The maximum number of blocks allowed in each phase was 8, at which point training progressed seamlessly to the next phase. Each test block was repeated twice. The entire experiment took approximately 20 minutes or less.

All trials progressed in a seamless series. There were no pauses or markers between phases. When a new epoch began, the novel symptoms appeared. The response prompt for each trial indicated whether the learner was to guess the correct specialist (in epoch 1) or the correct diagnosis(in epochs 2 and 3).

Stimuli

Figures 11.2 and 11.3 show screen shots of the stimuli. Figure 11.2 shows a prompt and feedback screen from the first epoch, in which the correct specialist is explicitly and directly trained. The prompt asks the learner to press one of the keys 1, 2, or 3, and the feedback states the correct specialist and highlights the correct specialist with a heavy outline. The correct diagnosis is also indicated by a white (instead of grey) background. Figure 11.3 shows a prompt and feedback screen from the second and third epochs, in which the correct specialist is only indirectly trained via the diagnosis. Notice that the correct specialist is not indicated; only the correct diagnosis is shown. (It is only a random coincidence that Figures 2 and 3 both show Specialist 3 without a diagnosis. Across trials, the 'missing' specialist was counterbalanced, as indicated in the design of Table 11.1.)

Results

Participants

Participants volunteered for partial credit in introductory psychology courses at Indiana University. This subject pool has a median age of approximately 19 years, and is about 50–60% female. Procedures for protection of human subjects were approved by the local Institutional Review Board. There were 188 participants.

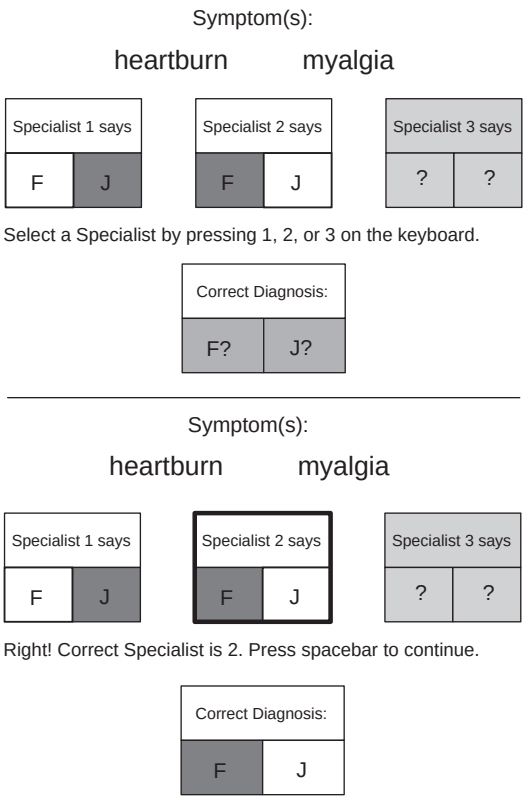

Fig. 11.2 For Epoch 1, involving explicit feedback about the correct specialist. *Upper*: Computer display for cues with response prompt. *Lower*: Computer display for cues with corrective feedback, after the correct response was made.

Learning criterion

For purposes of data analysis, epochs were excluded if accuracy did not achieve at least 58% in both training phases by the final block or training. The criterion was selected arbitrarily as a compromise between excluding too many subjects and including too many poor learners. Results did not change in any qualitative way with different criteria. The criterion resulted in 180, 122, and 131 subjects (out of 188) included in each of epochs 1, 2, and 3, respectively. Many or most of the excluded subjects appeared to have been unmotivated to learn, as their response times were on the order of 200 ms. or less, which indicates pressing a key as quickly as a stimulus appears without processing its attributes.

Choice in the test phase

Table 11.2 shows the choice preferences in the test phase, collapsed across participants. All three epochs show evidence of blocking, i.e., for the BC or BD tests, the response tends to be consistent with C or D more than with B.

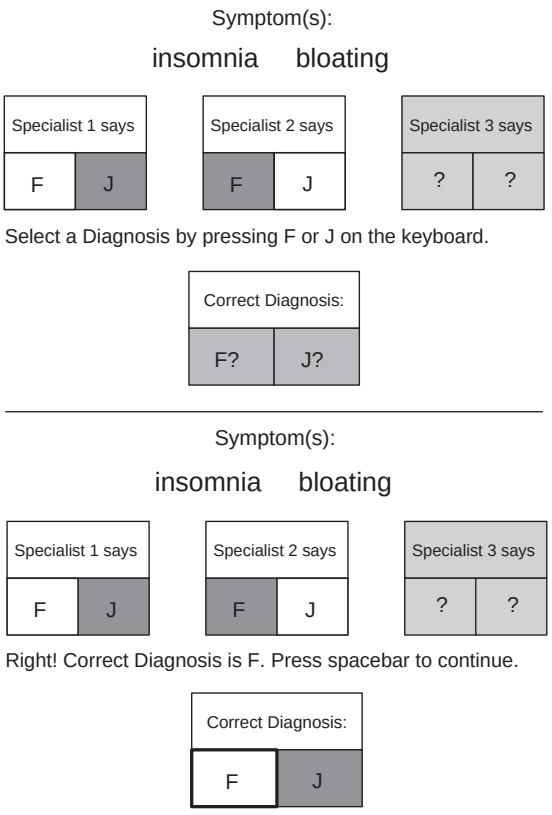

Fig. 11.3 For Epochs 2 and 3, involving feedback only about the correct diagnosis. *Upper*: Computer display for cues with response prompt. *Lower*: Computer display for cues with corrective feedback, after the correct response was made.

Table 11.2 Test phase response percentages, collapsed across subjects

	Response consistent with A/B or C/D					
	Epoch 1 Backward Blocking Explicit Specialist		**Epoch 2 Forward Blocking Indirect Feedback**		**Epoch 3 Backward Blocking Indirect Feedback**	
Test cues	A/B	C/D	A/B	C/D	A/B	C/D
AC or AD	54.5	45.5	59.8	40.2	52.3	47.8
BC or BD	42.1	57.9	44.5	55.5	45.3	54.7

The magnitude of the preference is weak, however. For example, the magnitude of backward blocking in Epoch 1 is notably weaker than that reported by Kruschke and Blair (2000).

We can only speculate as to why the blocking effect is so weak, but presumably it is because of the complex stimulus display and distraction by the disease diagnoses that were irrelevant in Epoch 1. Subsequent epochs also show weak

magnitudes of blocking, presumably because of the difficulty of inferring the correct specialist indirectly from the diagnosis. We will say more about the magnitude of blocking in the discussion after the statistical analysis.

Bayesian statistical analysis

The data were analyzed using Bayesian methods. In a Bayesian analysis, a descriptive model of the data is defined, and the parameter values of the model are estimated. The Bayesian analysis yields an entire posterior distribution of parameter values, not merely a single best-fitting parameter value. One reason to prefer Bayesian methods over traditional null hypothesis significance testing (NHST) is that the Bayesian analysis yields an explicit distribution regarding the believability of various underlying choice probabilities, given the experiment data. Another reason to prefer a Bayesian approach is that individual differences are explicitly modelled and taken into account. In the following paragraphs, the model is first defined, followed by a description of how the posterior distribution was generated, followed, finally, by a description of the posterior distribution itself.

In the test phase of any epoch, each participant saw the two context types (i.e., either AC/AD or BC/BD) eight times, because there were two repetitions of the test block in Table 1. For each test type, the i^{th} participant's 8 responses to that type were modeled as a random sample from a binomial distribution having underlying probability $\theta_{A/B,i}$ of selecting the response consistent with the A or B cue, and probability $\theta_{C/D,i} = 1 - \theta_{A/B,i}$ of selecting the response consistent with the C or D cue.

The individuals' probabilities, $\theta_{C/D,i}$, were modeled as a random selection from an overarching beta distribution that had (1) a parameter $\mu_{C/D}$ that specifies the central tendency of the group, and (2) a parameter κ that specifies how tightly the individuals were clustered around that central tendency. (In detail, the two 'shape parameters' of the beta distribution for $\mu_{C/D}$ were $a = \mu_{C/D}\kappa + 1$ and $b = (1 - \mu_{C/D})\kappa + 1$.)

The primary goal of the analysis is to generate a posterior estimate of the $\mu_{C/D}$ parameter for each test type. The $\mu_{C/D}$ parameter represents the overall response propensity for the type of context. The value of $\mu_{C/D}$ can range between 0 and 1, and can be thought of as the underlying probability of responding consistently with the C/D cues. When $\mu_{C/D} > .5$, the C/D cues are dominating the competing cues. When $\mu_{C/D} < .5$, the C/D cues are being dominated by the competing cues. If there is blocking, then the analysis should show that the credible values of $\mu_{C/D}$ are greater than 0.5, when the competing cue is B.

The hyperprior on $\mu_{C/D}$ was a uniform on the interval (0,1). This means that the prior was very noncommittal and gave the full range of $\mu_{C/D}$ values equal credibility. The hyperprior on κ was a gamma density with shape and rate parameter values of 0.01 (censored at 0.3 so that the random samples in the MCMC chain did not cause overflow errors in the beta density). This again means that the prior on κ was very noncommittal, allowing a huge range of possibilities, but emphasizing small values of κ that reflect large individual differences and a conservative estimate of $\mu_{C/D}$. The posterior distributions were robust to reasonable changes in the diffuse hyperpriors.

This hierarchical model allowed individual differences to be captured by variation in participant-level binomial probabilities, which in turn were mutually informed by being modeled as representative samples from the same higher level beta distribution. The higher level beta distribution captures across-subject response tendencies for each context type. The posterior certainty in the beta parameters depends on the consistency of response tendencies across subjects.

There is no general analytical solution for deriving the forms of the posterior distributions in hierarchical models. Nevertheless, the posteriors can be accurately estimated by generating large representative samples. The large samples include parameter values that are consistent with the data and the prior. The samples are generated by taking a random walk through the high-dimensional parameter space. Each step in the walk lands on a point for which the combination of parameter values is credible, given the data. Thus, after a large number of steps in the random walk, the sampled points provide an accurate reflection of the underlying continuous posterior distribution. The distribution of points also inherently reveals any correlations among credible parameter values.

The posterior distribution was determined by Markov chain Monte Carlo (MCMC) approximation. The program for generating the sample was written in the R language (Ihaka and Gentleman, 1996), using the BRugs interface (Thomas, 2004) to the OpenBUGS version (Thomas, O'Hara, Ligges, and Sturtz, 2006) of BUGS (Gilks,Thomas, and Spiegelhalter, 1994). Three parallel MCMC chains were simulated, using a burn-in of 10,000 steps and thinning of 200 steps. This burn-in and thinning produced well-mixed chains with small auto-correlation, so the posterior sample is very trustworthy. From each of the three chains, 1,000 steps were retained to represent the posterior, yielding 3,000 representative parameter values.

Figure 11.4 shows histograms of the believable values of the $\mu_{C/D}$ parameters. The upper three histograms indicate results from Epoch 1, i.e. backward

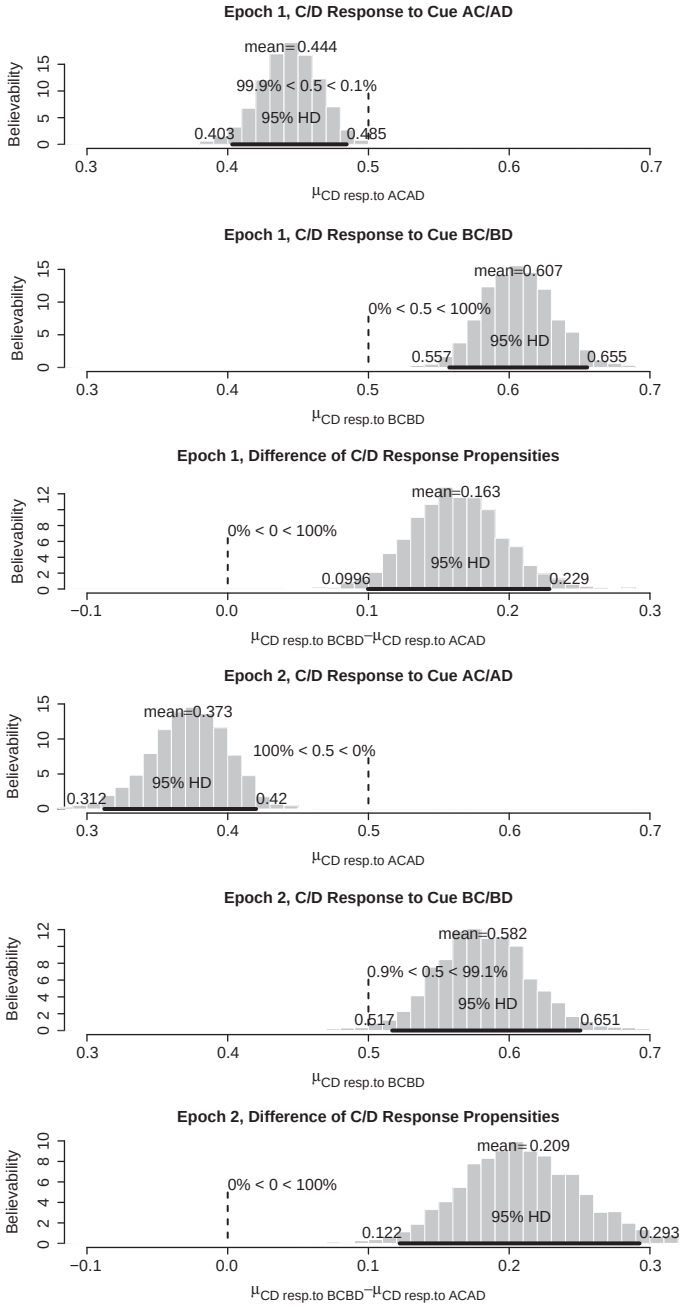

Fig. 11.4 Posterior distribution of C/D preference for the group. *Upper three histograms*: Epoch 1; backward blocking with direct feedback. *Middle three histograms*: Epoch 2; forward blocking with indirect feedback. *Lower three histograms (next page)*: Epoch 3; backward blocking with indirect feedback. The dark bar labeled '95% HD' spans the 95% highest density interval, such that all parameter values within the interval have higher believability than values outside the interval, and the interval covers 95% of the believable values.

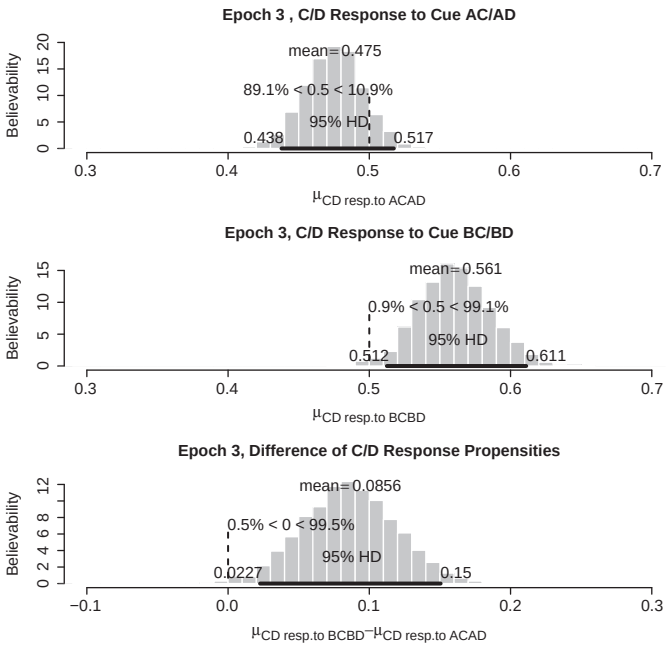

Fig. 11.4 (continued)

blocking with direct feedback regarding the relevant specialist. The upper histogram indicates that for the test probes AC or AD, the credible values of the response propensity are virtually all less than .5, i.e., people prefer the response consistent with cue A. The second histogram of the set indicates that for test probes BC or BD, the credible response propensities are all well above 0.5, indicating a robust backward blocking effect. The third histogram in the set indicates that the difference between the two types of probes is credibly different from zero.

The histograms in Figure 11.4 for Epoch 2 indicate that there was credible forward blocking of cues to relevance, even when there was only indirect feedback regarding the correspondence of context cues to response cues. In particular, virtually all the believable response propensities to BC or BD cues are in favor of the C/D consistent response.

Most important for our present purposes, the lower set of histograms in Figure 11.4 for Epoch 3 indicates that there was credible *backward* blocking of cues to relevance, even when there was only indirect feedback regarding the correspondence of context cues to response cues. Specifically, the distribution that estimates the C/D propensities for cues BC or BD falls mostly (99.1%)

above 0.5. The lowest histogram shows that the C/D propensity for BC or BD tests is larger than the C/D propensity for AC or AD tests, with nearly all the distribution falling above zero.

Summary and discussion of experiment results

The Bayesian analysis of the data incorporated a model of individual differences and yielded an explicit representation of credible response propensities. The analysis revealed that it is highly credible that there is forward and backward blocking of cues to relevance.

For devotees of the twentieth century ritual of null hypothesis significance testing, a chi-square analysis is hereby provided. In Table 11.2, for each epoch's 2×2 table, a chi-square test of independence was conducted on the raw frequencies. These three tests correspond to the lowest histogram in each of the three panels of Figure 11.4. For Epoch 1, $\chi^2(df = 1, N = 2880) = 44.55, p < .001$; for Epoch 2, $\chi^2(df = 1, N = 1952) = 46.19, p < .001$; for Epoch 3, $\chi^2(df = 1, N = 2096) = 9.90, p < .002$. A chi-square test can also be conducted on the the BD trials alone, i.e., the lower row of Table 11.2. These three tests correspond to the middle histogram in each of the three panels of Figure 11.4. For Epoch 1, $\chi^2(df = 1, N = 1440) = 36.10, p < .001$; for Epoch 2, $\chi^2(df = 1, N = 976) = 11.95, p < .001$; for Epoch 3, $\chi^2(df = 1, N = 1048) = 9.16, p < .003$. The implication from these tests is that there was highly significant backward blocking in all three Epochs. These standard analyses assume that all individuals have the same underlying magnitude of response preference. This is not a reasonable assumption; the Bayesian analysis does not make it. Furthermore, the standard computation of p values assumes that N was fixed in advance, and data collection was stopped when that N was reached. In fact, the data were collected for a set number of weeks during which participants volunteered, and data were excluded if learners did not reach the accuracy criterion. Bayesian analysis has no reliance on why data collection stopped. Notice also that the Bayesian analysis provides a complete distribution for credible values of the parameters (in Figure 11.4), while the chi-square analysis provides no interval estimate.

As mentioned earlier, the magnitudes of the forward and backward blocking effects were weak. We speculate that the small effects occurred because processing the corrective feedback was effortful and interfered with attentional re-allocation. For example, when the corrective feedback indicates disease F, the learner must first determine which specialist is consistent with that outcome, before then being able to determine which symptoms are relevant to selecting that specialist. If processing of the feedback interferes with allocation

of attention to symptoms, and if blocking of symptoms depends in part on attentional allocation, then the extra processing of the feedback may impair learned inattention to the blocked cue.

The limited variety of test trials in this particular experiment admits a different explanation of the results. In this alternative explanation, there is no blocking during learning, but there is instead a response bias at test. On tests involving BC or BD cues, the preference for C/D-consistent responses is a result of a response bias to choose specialist S2 whenever two context cues appear. The response bias is an inference from the training phases, for which it is the case that only two-cue contexts occurred when specialist S2 was relevant. This general bias is overcome on tests involving AC or AD cues, because the association from A to S1 is very strong. Only future experiments will be able to distinguish the two explanations, perhaps by including tests involving single-cue contexts. For example, the two test items $B_{S1}\,S1_F\,S2_J$ and $D_{S2}\,S1_F\,S2_J$ involve a single context cue, and therefore would not suffer the hypothesized two-context-cue bias toward specialist S2, but blocking would predict a difference in response preferences across the two items. Until follow-up experiments are conducted, we must rest with the argument that structurally analogous previous experiments, involving blocking of cues to outcomes instead of cues to relevance, have shown less ambiguous blocking phenomena (e.g., Kruschke and Blair, 2000).

The current results at least indicate proof of concept: In principle, it is possible for people to learn which context cues indicate which response cues to attend to. And, most importantly, we have demonstrated data consistent with backward blocking of cues to relevance. Future experiments will attempt to use context and response cues that are not so blatantly distinct as symptoms and medical specialists. With enough training, people should be able to learn in those situations too, and presumably also show backward blocking.

Modelling

The experiment of the previous section showed results consistent with backward blocking of contextual cues to relevance. In this section we show that the behavior can be mimicked by a locally Bayesian learning model in which the first layer learns to allocate attention and the second layer learns to generate outcomes.

The architecture of the model is illustrated in Figure 11.5. Each node in the lower layer represents a cue that can be present or absent. Notice that the lower layer consists of 11 cues, including the symptoms and the specialist information. The symptoms and specialists were presented in the experiment

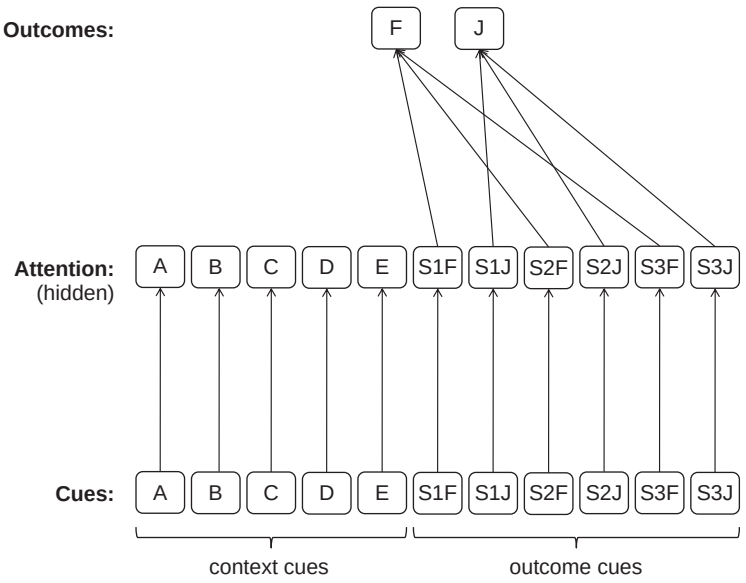

Fig. 11.5 Model architecture. Arrows denote the most credible associations at the beginning of training (i.e., the mean of the prior distribution). The cues are marked at the bottom as context or outcome cues, but this marking is purely for the benefit of the reader, as the model does not 'know' which cues are context cues and which cues are outcome cues.

as different types of cues, but in principle they are both just present/absent bits of information. For example, when stimulus A_{S1} $S1_F$ $S2_J$ is presented, input nodes A, S1F, and S2J are activated.

The middle layer, otherwise called the hidden layer, represents attentionally gated cue activations. Each input cue has a corresponding node in the hidden layer. By default, each cue calls attention to itself, indicated in Figure 11.5 by the 1-to-1 arrows. Learning can cause the magnitude of the 1-to-1 connection to change, and also generate 'lateral' connections between input cues and attentional nodes. In particular, when a cue is blocked, the 1-to-1 connection from that cue to its own attentional gate may be reduced, and the lateral connection to its attentional gate from the blocking cue may become negative. More accurately described, one would say that lower values of the 1-to-1 connection become more believable, and negative values from the blocking cue to the blocked attention node become more believable. [Griffiths and LePelley (2009) have shown that *forward* blocking of *response* cues is unlikely to produce strong negative lateral connections, in at least some situations. It is not yet known

whether comparable experiments with backward blocking would lead to analogous conclusions. The present model can accommodate these results, at least in principle, by modulating the relative learnabilities of lateral and 1-to-1 weights.]

The upper layer represents outcomes, which are disease diagnoses in the present application. The associations from attentionally gated cues to outcomes must be learned. In principle, any of the cues could be indicative of any outcome. In the present application, only the cues corresponding to medical specialists happen to be correlated with the correct diagnoses. Moreover, the experiment presented the specialist cues in such a way that the corresponding diagnosis was explicitly indicated. In other words, the associative links from specialists to diagnoses were already suggested, and therefore the model has these links built in, as shown in Figure 11.5 by the arrows fanning into the outcome nodes. Through learning, these links can be altered, and other links from cues to outcomes can be established.

In locally Bayesian learning, multiple hypotheses about the associative weights are maintained, for both layers of weights. Learning consists of shifting credibility from unlikely hypothetical weights to likely hypothetical weights. The arrows in Figure 5 indicate the hypothetical weights with highest credibility at the beginning of the experiment, before training.

Intuitively, locally Bayesian learning proceeds in Epoch 3 as follows. The first phase, involving redundant-context cues, has cases of $A_{S1}\ B_{S1}\ S1_F\ S2_J \rightarrow F$, among others (see Table 11.1). This phase causes credibility to be enhanced on positive connections from cues A and B to attention on S1F and S1J, and perhaps also causes credibility to be enhanced on negative connections from cues A and B to attention on S2F and S2J. In other words, the model learns that when symptoms A or B are present, attend to specialist 1 and ignore specialist 2. Because the lower layer is a Bayesian system that keeps track of multiple candidate weight combinations, the system 'knows' that the training cases could be explained by either (1) A indicating S1 and B being irrelevant, or (2) B indicating S1 and A being irrelevant, or (3) A or B indicating S1. All of these plausible weight combinations retain some credibility. In the second phase, involving single-context cues, there are cases of $A_{S1}\ S1_F\ S2_J \rightarrow F$, among others (see Table 11.1). Of the weight combinations that remained credible after the first phase, the ones involving cue A gain increased credibility, thereby reducing the credibility of the ones involving cue B alone. In other words, locally Bayesian learning on the lower layer can account for backward blocking of contextual cues to relevance.

The remainder of this section is devoted to a detailed description of a particular instantiation of this approach. In this particular instantiation, the

outcome and attention nodes are modeled as individual Kalman filters. Kalman filters have been previously used by Dayan and Kakade (2001) to model backward blocking, and the idea of using layers of locally learning Kalman filters was mentioned by Kruschke (2006a, 2006b). We do not intend to argue that Kalman filters are the best way to instantiate the components of locally Bayesian learning. Indeed, Kalman filters can only represent linear mappings, and are therefore quite limiting. We use Kalman filters merely because they are convenient for computational tractability. Presumably the model structure used by Kruschke (2006b), which did not involve Kalman filters, would show similar qualitative behavior.

One goal of the detailed modeling is to demonstrate by computer simulation that the intuitive argument provided in the previous paragraphs is actually correct. A second goal of the reporting the mathematical details is to present novel derivations that have not been previously presented elsewhere. There is not sufficient space here to provide an extensive tutorial on Kalman filters, but a tutorial regarding Kalman filters applied to associative learning has been previously provided by Kruschke (2008). The reader, who is not concerned at this time with the mathematical implementation of the model, is invited to skip ahead to the next subsection where the model results are reported.

Formal description of layers of Kalman filters

The Kalman filter assumes that the output to be predicted is a scalar metric value y. In our experiment, the outcome is dichotomous (present/absent), which is represented as $y = 1$ or $y = 0$. The Kalman filter assumes that y is a linear function of the input vector $\vec{x}$ (thought of as a column matrix), with normally distributed noise in the output. The variance of the noise is the scalar value v. The associative weights are denoted by the vector $\vec{w}$ (thought of as a column matrix). Formally, then, the probability of a value y is

$$p(y \mid \vec{x}, \vec{w}, v) = \frac{1}{\sqrt{v}(2\pi)^{1/2}} \exp\left(-.5 \frac{(y - \vec{w}^T \vec{x})^2}{v} \right) \tag{11.1}$$

The value of v is considered to be a known constant, fixed by the modeler. Equation 11.1 is the likelihood function that specifies the probabilistic output of any node in the model. The normal distribution in Equation 1 is also denoted $N(y \mid \vec{w}^T \vec{x}, v)$.

Each outcome and attention node is a distinct Kalman filter. For example, the node for outcome F is a Kalman filter for which $y = 1$ means F is present, and $y = 0$ means F is absent. The input to that outcome node is the vector of attentional values from the hidden layer. The attentional values are the (means

of the) output values of the attentional Kalman filter nodes. Each attentional node is also a Kalman filter. The input to each attentional node is the vector of cue values.

In each Kalman filter, the value of $\vec{w}$ has uncertainty, meaning that each possible combination of weights has a degree of belief, and beliefs are spread over a wide range of weight combinations. When the model learns, belief is shifted toward weight combinations that are consistent with the training items. The initial state of the network has beliefs spread out symmetrically and diffusely over a wide range of weight combinations. The initial state is called the prior distribution. It is assumed to be normal with mean $\vec{\mu}$ (a column matrix) and covariance matrix **C**. Thus, the prior on $\vec{w}$ is

$$p(\vec{w} \mid \vec{\mu}, \mathbf{C}) = \frac{1}{\sqrt{|\mathbf{C}|}(2\pi)^{d/2}} \exp\left(-.5(\vec{w} - \vec{\mu})^T \mathbf{C}^{-1}(\vec{w} - \vec{\mu})\right) \tag{11.2}$$

where $|\mathbf{C}|$ is the determinant of **C**, and d is the dimensionality of $\vec{w}$, i.e., the number of input nodes. Equation 11.2 is merely the well-known formula for the multivariate normal distribution, and Equation 11.1 is merely the special case of that formula when $d = 1$ and $\mu = \vec{w}^T \vec{x}$. The normal distribution in Equation 11.2 is also denoted $N(\vec{w} \mid \vec{\mu}, \mathbf{C})$.

The mean vector for the weights was set to all zeros except for specific components that represented initial correspondences. The mean vector on weights fanning in to the outcome nodes was initialized to represent the explicit mappings indicated by the specialists. Therefore the mean weight to F from S1F was set to 1, as was the weight to F from S2F, and to J from S1J, and so forth. These mean connection weights of 1 are represented by the arrows in Figure 11.5. The mean vector on weights fanning in to the attention nodes was initialized to represent the 1-to-1 correspondence of cues to attention, but also to take into account only partial attention to any given cue. Therefore the 1-to-1 connections were initialized arbitrarily at a mean value of 0.1. These 1-to-1 connections are represented by the arrows in Figure 11.5.

The prior covariance matrix for each node specifies the prior certainty in the mean values. Because the specialists gave explicit diagnoses, the mean weights from specialists to outcome nodes had very small variances, i.e., very high certainties. But the weights to the outcome nodes from the context attention nodes had prior variances that were large, to reflect great initial uncertainty. The prior covariance matrices for the attention nodes were likewise set with small variances on the specialist cues and large variances on the context cues. This was because of the prior assumption that specialists do not inform each other, but the symptoms handled by each specialist must be learned.

Prediction by a Kalman-filter node

In propagating activation up a succession of Kalman filter nodes, the input to the upper layer is the mean output of the lower layer. The mean output of a Kalman filter is simply

$$\begin{aligned}\bar{y} &= \int d\vec{w}\, p(\vec{w} \mid \vec{\mu}, \mathbf{C}) \int dy\; y\, p(y \mid \vec{x}, \vec{w}, v) \\ &= \int d\vec{w}\, p(\vec{w} \mid \vec{\mu}, \mathbf{C})\, \vec{w}^T \vec{x} \\ &= \left(\int d\vec{w}\, p(\vec{w} \mid \vec{\mu}, \mathbf{C})\, \vec{w} \right)^T \vec{x} \\ &= \vec{\mu}^T \vec{x}\end{aligned} \tag{11.3}$$

In the architecture we use for expressing attentional gating, the attentional Kalman filter acts as a multiplier on the input cue. Formally, the hidden layer activation that acts as the input to the outcome nodes is the mean activation of the attentional Kalman filters times the corresponding input cue activations: $x_i^{hid} = \bar{y}_i x_i^{cue}$ where $\bar{y}_i$ is the mean output of the *i*-th attentional Kalman filter, as determined from Equation 11.3.

Learning by a Kalman-filter node

The values of $\vec{\mu}$ and $\mathbf{C}$ serve as the prior distribution for the trial's Bayesian updating. On a given trial, the input vector is $\vec{x}$ and the correct output value, a.k.a. the target, is denoted t (a scalar). It turns out (Meinhold and Singpurwalla, 1983) that the Bayesian updating formula simplifies to the following expressions:

$$\vec{\mu}' = \vec{\mu} + \mathbf{C}\,\vec{x}\left[v + \vec{x}^T \mathbf{C} \vec{x}\right]^{-1} \left(t - \vec{x}^T \vec{\mu}\right) \tag{11.4}$$

$$\mathbf{C}' = \mathbf{C} - \mathbf{C}\,\vec{x}\left[v + \vec{x}^T \mathbf{C}\,\vec{x}\right]^{-1} \vec{x}^T \mathbf{C} \tag{11.5}$$

Thus, for any given outcome node or attention node, Equations 11.4 and 11.5 are applied to update its beliefs. The computational simplicity of these updating equations is was makes the Kalman filter appealing as an implementation of Bayesian learning. The target for the outcome nodes is explicitly indicated by corrective feedback, and the input to the outcome nodes is the pattern of attentional activation as defined in the previous paragraph. For the attentional nodes, the input is the cue activation, but the target values for the attentional nodes will be defined in the next subsection.

In summary, at the beginning of a trial, the weight vector is distributed as in Equation 11.2. Then target and input for the trial are provided, and beliefs

regarding credible weight combinations are adjusted by Bayes' rule. The posterior distribution of the weights conveniently turns out to be again normal, with mean and covariance given by Equations 11.4 and 11.5.[2]

Finding the attention that maximizes a desired output

To find a target for the attention nodes, we want to find the attentional input to the outcome nodes that would maximize the probability of the correct outcome. In other words, given the target outcome values, t_k of outcome k, we want to find the attention-node values $\vec{x}_t$ that maximize the joint probability of the outcome values:

$$\begin{aligned}\vec{x}_t &= \arg\max_{\vec{x}} \prod_k p\left(t_k \mid \vec{x}\right) \\ &= \arg\max_{\vec{x}} \prod_k \int d\vec{w}\; p(t_k \mid \vec{x},\vec{w},v)\, p(\vec{w} \mid \vec{\mu}, \mathbf{C})\end{aligned} \tag{11.6}$$

A formal identity regarding the product of Gaussians states that

$$\begin{aligned}&N(\vec{w}\,|\,\vec{\mu},\mathbf{C})\, N(\vec{x}^T\vec{w}\,|\,t,v) \\ &\quad = N(t\,|\,\vec{x}^T\vec{\mu},v+\vec{x}^T\mathbf{C}\vec{x})\, N(\vec{w}\,|\,\vec{m},\mathbf{V})\end{aligned}$$

where

$$\begin{aligned}\vec{m} &= \mathbf{V}\left(\mathbf{C}^{-1}\mu + \vec{x}\,t\,/\,v\right) \\ \mathbf{V} &= \left(\mathbf{C}^{-1} + \vec{x}\,\vec{x}^T/\,v\right)^{-1}\end{aligned}$$

[2] The full Kalman filter also assumes that the distribution of $\vec{w}$ changes in time, separately from and before any updating in beliefs inferred from observed data. This dynamic aspect of the weights is assumed to be linear at any given trial, so that the mean weights are dynamically changed into some linear transformation of the current mean weights. Dayan et al. (2000) used this mechanism to model unbiased diffusion of weights through time. Because afficianados of the Kalman filter may wonder what we did with the dynamic mechanism, it is summarized in this footnote. Denote the dynamic linear transformation by **D**. There is also assumed to be some additive change in the covariances of the weights; typically this is thought of as a constant increment in the uncertainty of the weights as time progresses. Denote the added uncertainty by **U**. Formally, then, at the beginning of each trial, the weight distribution is dynamically changed as follows: $\vec{\mu}^* = \mathbf{D}\vec{\mu}$ and $\mathbf{C}^* = \mathbf{DCD}^T + \mathbf{U}$. In all of our applications, we assume that **D** is the identity matrix and that **U** = 0. That is, we assume that the weights are not systematically changing through time. This restriction implies that all the behavior of the model comes from Bayesian learning, not from additional dynamic assumption.

Hence the integral in Equation 11.6 can be rewritten as

$$\begin{aligned}
&\int d\vec{w}\, p(t_k|\vec{x},\vec{w},v)\, p(\vec{w}\,|\,\vec{\mu},\mathbf{C}) \\
&= \int d\vec{w}\; N\left(t_k\,|\,\vec{x}^T\vec{w},v\right)N(\vec{w}\,|\,\vec{\mu},\mathbf{C}) \\
&= \int d\vec{w}\; N\left(t_k\,|\,\vec{x}^T\vec{\mu},v+\vec{x}^T\mathbf{C}\vec{x}\right)N(\vec{w}\,|\,\vec{m},\mathbf{V}) \qquad (11.7)\\
&= N\left(t_k\,|\,\vec{x}^T\vec{\mu},v+\vec{x}^T\mathbf{C}\vec{x}\right)\int d\vec{w}\, N(\vec{w}\,|\,\vec{m},\mathbf{V}) \\
&= N\left(t_k\,|\,\vec{x}^T\vec{\mu},v+\vec{x}^T\mathbf{C}\vec{x}\right)
\end{aligned}$$

We used numerical approximation(specifically a Newton-Raphson method applied to the derivative) to find the $\vec{x}$ that maximizes Equations 11.6 and 11.7.

Model result: Single-layer Bayesian model fails

The experiment was designed such that no context cue is individually correlated with the correct diagnosis. Inspection of Table 11.1 reveals that every context cue occurs as often with outcome *F* and with outcome *J*. Therefore, a single-layer Kalman filter is unable to learn the correct responses in either the single-context phase or the redundant-context phase.

Because complex models have an uncanny way of confounding intuitions, especially when applied to complex designs, we simulated a single-layer Kalman filter to be sure that such a model truly was unable to learn the mapping. The results verified that the model could only produce 50–50 (i.e., chance) responding for Epochs 2 and 3.

Model result: Locally Bayesian learning succeeds

To fit the output of the model to the human response proportions in Table 11.2, we had to map the model output, expressed as $\overline{y}_k$, to response proportions. We did this via the often used softmax rule: $p(K) = \exp(\gamma\overline{y}_K)\big/\sum_k \exp(\gamma\overline{y}_k)$, where $\gamma > 0$ is a parameter called the 'decisiveness' of the choice. When γ is large, small differences in the outcome activations lead to large differences in choice proportion, but when γ is small, outcome activations must be very different to produce much difference in choice proportion. We arbitrarily set $\gamma = 1$.

The only free parameter was the noise variable v in Equation 11.1. The noise parameter acts much like a learning rate, as can be seen by its appearance in the update Equations 11.4 and 11.5. When v is larger, learning is slower.

The model was trained in each phase on the median number of blocks that human learners took. For successive phases in the three consecutive epochs, the median number of blocks was 3, 2, 4, 2, 3, and 2, respectively. The average of 40 simulated subjects, each with a different permutation of trials within

Table 11.3 Locally Bayesian model behavior

	Response consistent with A/B or C/D					
	Epoch 1 Backward Blocking Explicit Specialist		Epoch 2 Forward Blocking Indirect Feedback		Epoch 3 Backward Blocking Indirect Feedback	
Test cues	A/B	C/D	A/B	C/D	A/B	C/D
AC or AD	56.1	43.9	56.2	43.8	56.0	44.0
BC or BD	44.2	55.8	44.4	55.6	44.3	55.7

blocks, was used as the model prediction. The fit of the model was measured simply as sum-squared-deviation between the human response percentages in Table 11.2 and the corresponding model response percentages.

The best fit used $v = 0.047$, and the average model output is shown in Table 11.3. The model shows forward and backward blocking very similar to the human preferences in Table 11.2. The qualitative trend of the model is quite robust against changes in parameter values. The main point of the simulation is to demonstrate that the locally Bayesian model does indeed produce backward blocking of cues to relevance, as shown by the model behavior in Epoch 3. The magnitude of this backward blocking can be made larger with other parameter values. The magnitude is small in this simulation specifically to match the small magnitude shown by human learners.

The fit by the model is not 'spot on' the data, but the model uses only one free parameter and there are many aspects of the human procedure not directly imitated in the model procedure. For instance, all simulated subjects used exactly the median number of training blocks, but many human subjects took more blocks to learn. All simulated subjects used the same value of V, but presumably different human learners had different learning rates. The simulated model had a constant value of v throughout all three epochs, but human learners may have 'learned to learn' as they became familiar with the stimulus arrangement and task during training. And, perhaps most importantly, the simulation started the 1-to-1 attention links at the same neutral prior for every epoch, but people may have had progressively more certain priors in successive epochs. The purpose of the model simulation is to demonstrate that even with these simplifications, the locally Bayesian model shows robust backward blocking of cues to relevance.

Conclusion

We have provided some data that suggest that people exhibit backward blocking of cues to relevance. One way of modeling this behavior is with layers of locally Bayesian learning. The lower layer learns to allocate attention to cues,

and the upper layer learns to produce outcomes based on the attended cues. Because there is Bayesian learning within each layer, the model can exhibit backward blocking of associations within each layer.

The novel contributions of this chapter are (1) the empirical suggestion of backward blocking of cues to relevance, and (2) the implementation of locally Bayesian learning as layers of Kalman filters. Backward blocking of cues to relevance may justify locally Bayesian learning in the attentional layer of Kruschke's (2006) model. The results of the new experiment reported here may have alternative interpretations, but we hope that this work may provoke interesting follow-up research to distinguish alternative accounts. Any competing model should also account for the spectrum of phenomena addressed by locally Bayesian learning applied to attentional learning (Kruschke, 2006b).

Acknowledgements

For help administering the experiments, thanks go to Kelsey Buckingham, Kevin Clemens, Alyssa Heggen, Kaitlyn Smith, Kari Vann, and Phaedra Willson. For providing pilot data and discussion, we thank Rima Hanania and Richard Hullinger. The editors, Chris Mitchell and Mike LePelley, provided helpful suggestions for clarification and discussion. Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to kruschke@indiana.edu. Supplementary information can be found at http://www.indiana.edu/~kruschke/

References

Chater, N., Tenenbaum, J.B., & Yuille,A.(Eds.). (2006, July). Special issue: Probabilistic models of cognition. *Trends in Cognitive Sciences*, *10*(7), 287–344.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. 13, pp. 451–457). Cambridge, MA: MIT Press. Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

De Houwer, J. & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *The Quartely Journal of Experimental Psychology*, *55B*, 289–310.

Deneve, S. (2008). Bayesian spiking neurons II: Learning. *Neural Computation*, *20*, 118–145.

Dickinson, A. (2001). Causal learning: Association versus computation. *Current Directions in Psychological Science*, *10*, 127–132.

Dickinson, A. & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, *49B*, 60–80.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*(1), 169–177.

Griffiths, O., & Le Pelley, M. E. (2009). Attentional changes in blocking are not a consequence of lateral inhibition. *Learning and Behavior*, *37*(1), 27–41.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 299–314. (Withother contributors listed at http://www.r-project.org/)

Kamin, L. J. (1968). 'Attention-like' processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–33). Coral Gables, FL: University of Miami Press.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812–863.

Kruschle, J. K. (2005). Learning involves attention. In G. Houghton (Ed.), *Connectionist models in cognitive psychology* (Ch 4, pp. 113–140). Hove, East Sussex: Psychology Press.

Kruschke, J. K. (2006a). Locally Bayesian learning. In Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 453–458).

Kruschke, J. K. (2006b). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*(4), 677–699.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior*, *36*(3), 210–226.

Kruschke, J. K. (2009). Highlighting: A canonical experiment. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 153–185). Academic Press.

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

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 830–845.

Le Pelley, M. E., Beesley, T., & Suret, M. B. (2007). Blocking of human causal learning involves learned changes in stimulus processing. *The Quarterly Journal of Experimental Psychology*, *60*(11), 1468–1476.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298.

Mackintosh, N. J., & Turner, C. (1971). Blocking as a function of novelty of CS and predictability of UCS. *Quarterly Journal of Experimental Psychology*, *23*, 359–366.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General*, *118*, 417–421.

Meinhold, R. J., & Singpurwalla, N. D. (1983). Understanding the Kalman filter. *American Statistician*, *37*(2), 123–127.

Miller, R. R. & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 22, pp. 51–92). San Diego, CA: Academic Press.

Mitchell, C. J., Harris, J. A., Westbrook, R. F., & Griffiths, O. (2008). Changes in cue associability across training in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*(4), 423–436.

Nelson, J. B. (2002). Context specificity of excitation and inhibition in ambigous stimuli. *Learning and Motivation*, *33*, 284–310.

Rosas, J. M., Callejas-Aguilera, J. E., Ramos-Alvarez, M. M., & Abad, M. J. F. (2006). Revision of retrieval theory of forgetting: What does make information context-specific? *International Journal of Psychology and Psychological Therapy*, *6*(2), 147–166.

Shanks, D.R. (1985). Forward and backward blocking inhuman contingency judgement. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, *37B*, 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.

Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 193–204.

Tenenbaum, J. B. & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun, and K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 35–42). Cambridge, MA: MIT Press.

Thomas, A. (2004). *BRugs user manual (the R interface to BUGS).* http://mathstat.helsinki.fi/openbugs/data/Docu/ BRugs%20Manual.html.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006, March). Making BUGS open. *R News*, *6*(1), 12–17.

Van Hamme,L.J.& Wasserman, E.A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*(3), 127–151.

Wills, A. J., Lavric, A., Croft, G. S., & Hodgson, T. L. (2007). Predictive learning, prediction errors, and attention: Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience*, *19*(5), 843–854.

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yang, L.-X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Learning, Memory*, *29*(4), 663–679.