CHAPTER FIVE

Highlighting: A Canonical Experiment

John K. Kruschke[1]

Contents

Abstract

Highlighting is a perplexing effect in learning, in which shared features are more strongly associated with early learned outcomes but distinctive features are more strongly associated with later learned outcomes. The effect has been widely observed with different stimuli, procedures, and application domains. It continues to discomfit many theories of learning. This chapter provides results from a "canonical" design in which the base rates of early and late outcomes are equalized. This balanced design yields data that pose a challenge to models that have relied on differential base rates of past designs to mimic highlighting. The data are available at the author's Web site as a test bed for models. A Bayesian data analysis is also reported that provides explicit posterior distributions over choice probabilities. The posterior distribution is also available online.

[1] The author's World Wide Web page is at http://www.indiana.edu/~kruschke/.

Psychology of Learning and Motivation, Volume 51
ISSN 0079-7421, DOI: 10.1016/S0079-7421(09)51005-5

1. Cue-Outcome Learning as a Window on Cognition

It is easy for a person to learn that when "ocean" and "arrow" appear on a computer screen, s/he should press the "F" key, but when "ocean" and "tulip" appear on the screen, s/he should press the "J" key. In a standard learning procedure, a person sees the cue words, presses the key s/he thinks is correct, and receives the correct outcome. After many repetitions, accuracy improves. This procedure and the finding that people are able to learn are, in a word, dull.

One fact that makes the exercise a little less dull is that cue-outcome learning in the lab is a distillation of a type of high-stakes learning that happens in real life. As examples: Physicians learn which symptoms indicate deadly diseases, and stock brokers learn which financial markers indicate times to buy or sell millions. If this sort of learning is to be understood by cognitive scientists, they need to study it in simplified and controlled laboratory experiments. Lab experiments can rarely impose consequences such as bankruptcy or death, however. The innocuous and bland laboratory procedures may therefore be described, in polite company, as less than scintillating.

What elevates cue-outcome learning from the banal to the fascinating is that people may learn, and respond to novel cues, in ways that are perplexing, if not downright bizarre. Although learners have blazed mental pathways from cues to correct outcomes, those pathways may be so convoluted that it is puzzling how humankind has blundered its way to the top of the food chain.

This chapter focuses on one puzzling phenomenon in cue-outcome learning, called *highlighting*. It is interesting because it violates (many) prescriptions for what a rational learner *should* do, and it is interesting because it deviates from what (many) learning theories *can* do. Highlighting is vexing because it crashes the parties of many established learning paradigms, when propriety would prefer to ignore it. But highlighting is also revealing, forcing theorists to find mechanisms that can explain it. Once revealed, the mechanisms may be seen to be fundamental aspects of learning, not just bad behavior.

After a brief review of previous work on highlighting, the primary goal of this chapter is to report new data from a "canonical" highlighting experiment. It is hoped that these data can serve as a test bed for models of learning. The data are available on the author's Web page, and so is the computer program for the experiment itself. The chapter also provides a novel Bayesian analysis of the data, unlike previous reports. The Bayesian analysis yields distributions of believable response propensities. These posterior distributions are also available on the author's Web page.

The chapter concludes with a brief discussion of the continuing challenges posed by highlighting for recent models of learning, including Bayesian learning models.

2. Highlighting

2.1. The Phenomenon

Suppose a person initially learns that when "ocean" and "arrow" appear on a computer screen, s/he should press the "F" key. The person subsequently learns that when "ocean" and "tulip" appear on the screen, s/he should press the "J" key. Notice that "arrow" is a perfect predictor of "F," and "tulip" is a perfect predictor of "J," whereas "ocean" is an imperfect predictor. Thus, there is a symmetry between the two responses, each having a unique perfect predictor, and sharing an imperfect predictor. Given this simple symmetry, it is reasonable to assume that the person learned the symmetry. This assumption can be assayed by testing the person with the single cue word "ocean." If the cue has been appropriately learned to be an equally imperfect predictor of the two outcomes, then the person should respond equally with the outcomes. Across many learners, however, there is a strong tendency to prefer the early learned "F" outcome. Unfortunately for learning theorists, this preference cannot be trivially explained as a generic primacy bias in response to ambiguous cues, because when learners are presented with the ambiguous cue combination "arrow" and "tulip," there is a strong preference for the later learned "J" outcome.

The phenomenon occurs for a variety of cues and outcomes and is not restricted to cues as words and responses as letters. Therefore, the cue-outcome structure is here redescribed with generic notation, abstracted from any irrelevant concrete instantiation. Early in training, the learner experiences cases of cues PE and I together indicating outcome E. This case is denoted I.PE → E. This case is trained until the learner knows it well. Then the learner is trained with cases of I.PL → L, in which cue PL and outcome L have not been previously trained. Notice that the cue structure is symmetric: Each outcome has a single perfectly predictive cue and the outcomes share the cue I. The only difference is that outcome E is trained *early*, and outcome L is trained *late*. Thus, cue PE is a *perfect* predictor of the *early* outcome, and cue PL is a *perfect* predictor of the *later* outcome, while cue I is an *imperfect* predictor. Interspersed training of I.PE → E and I.PL → L continues until both are learned well. Near-perfect accuracy is not difficult to attain. After training, when probed with cue I by itself, people are not impartial, instead strongly preferring outcome E. On the other hand, when presented with the cue pair PE.PL, people strongly prefer outcome L.

This torsion in people's preferences, going one way for I but twisting the opposite way for PE.PL, is called the "highlighting" effect. The appellation derives from two sources. First, highlighting refers to a theoretical interpretation of the empirical effect. In this interpretation, cue PL is attentionally highlighted during the learning of the cases I.PL → L. When experiencing I.PL → L, learners shift attention away from cue I, which is already associated with outcome E, toward cue PL. This theory will be explained more thoroughly later. The second motivation for the name "highlighting" is to juxtapose the empirical finding as complementary to the classic "blocking" phenomenon in associative learning (Kamin, 1968; Shanks, 1985), which can be at least partially explained by learned *inattention* to a cue, as opposed to learned highlighting of a cue (Kruschke, 2003b; Kruschke & Blair, 2000). Blocking will also be described in more detail later.

2.2. Highlighting Discomfits Theories of Learning

The highlighting effect is curious because people appear to have learned an asymmetrical cue-outcome structure despite the simple symmetry in the environment. What makes the phenomenon deeply interesting, however, is that most theories of learning cannot explain it.

Simple associative theories such as the Rescorla–Wagner model (Rescorla & Wagner, 1972) predict that after sufficient training, the associative weights reach asymptotic values that are symmetric. This symmetry emerges over several trials of later intermixed training. The initially learned association from I to E is reduced by subsequent cases of I.PL → L, because I has thereby occurred without E. The initially moderate association from PE to E increases when cases of I.PE → E recur, because cue I no longer predicts E very strongly. Eventually, the Rescorla–Wagner model accurately learns the symmetry, unlike people, who persist in the asymmetry even after fairly extended training. (Markman (1989) provides an alternative proof.)

Associative models that adjust cue salience or learning rates according to the novelty of the cues also fail to capture the effect. For example, a model was proposed by Shanks (1992) in which the salience of each cue is inversely proportional to a running estimate of its base rate. In other words, rare cues are more salient than frequent cues. While it is quite plausible that some form of novelty salience is at work in learning, and no doubt some phenomena do demand such a mechanism for adequate explanation, the particular mechanism in the proposed model does not account for effects closely related to highlighting (Kruschke, 1996). The novelty-salience model has not yet been fit to the new data reported in this chapter, but the model would probably have difficulty because, in the new experiment, what is initially a rare cue becomes a frequent cue, and *vice versa*.

Other variations of associative models set the learning rates of expected-but-absent cues to negative values (Markman, 1989; Tassoni, 1995; Van

Hamme & Wasserman, 1994). It is plausible that absent-but-expected cues are represented as explicitly absent in human learning, and perhaps some phenomena do demand such a representation for adequate explanation. But highlighting is not accounted for by these models. One difficulty with some of these models is that they do not propose a specific mechanism by which cue expectations are learned. Even when such mechanisms are specified, the new data presented in this chapter pose challenges for the models, because the long-run symmetry of the design (to be described later) implies that the absence of PE in I.PL trials may trade off symmetrically with the absence of PL in I.PE trials.

Various Bayesian models of learning fail to capture the effect. Several Bayesian models, such as the rational model (Anderson, 1990, 1991), the Kalman filter (Dayan, Kakade, & Montague, 2000), and sigmoid-belief networks (Courville, Daw, Gordon, & Touretzky, 2004), assume that all instances, regardless of their time of occurrence, are equally representative of the underlying cue-outcome association. In other words, the models are not sensitive to trial order, and, in particular, they cannot show highlighting (Daw, Courville, & Dayan, 2008; Kruschke, 2006b, 2006c).[1] These models' insensitivity to trial order is not a necessary shortcoming of all Bayesian models, however. These particular models ignore time (or trial) merely as a convenient mathematical simplification. Future Bayesian models might explicitly incorporate temporal dependencies.

Another approach is to try to explain highlighting as an inference during responding at test, rather than as an asymmetry during learning. The eliminative inference model (ELMO; Juslin, Wennerholm, & Winman, 2001; Winman, Wennerholm, & Juslin, 2003) is based on the idea that the test probe PE.PL is recognized to be an *un*known cue combination, and therefore known outcomes can be eliminated. If outcome E is well known, but outcome L is not, then E is eliminated and response L is preferred. Test probe I, on the other hand, is similar enough to learned rules that the known outcome E is evoked. There is good evidence that people do use some form of eliminative inference in some situations (Juslin et al.; Kruschke & Bradley, 1995). Unfortunately, it cannot account for highlighting. In particular, eliminative inference does not apply when all outcomes are well learned, but highlighting still occurs robustly in human preferences. Various details of response preferences are not captured by the ELMO model (Kruschke, 2001b, 2003a).

[1] The Kalman filter has a dynamic process that is sensitive to trial order, but the published versions of this mechanism do not account for highlighting (Daw et al., 2008; Kruschke, 2006b). And unlike the associative weights, the dynamic process parameters in the standard Kalman filter do not learn from training, but are fixed in advance. The rational model (Anderson, 1990, 1991) uses approximations that produce trial-order sensitivities, but these do not mimic highlighting (Kruschke, 2006b, 2006c).

The point of this section is merely to claim that the highlighting phenomenon is truly perplexing for many models of learning. There is not space here to thoroughly review all the contending models and the data that disconfirm them. The various references cited above provide many gory details of models impaled upon spikes of data.

2.2.1. Highlighting Is Explained by Attention Shifting

If all those theories do not explain highlighting, what does? A key insight was provided by Medin and Edelson (1988) (see also Medin & Bettger, 1991), who suggested that when learning I.PE → E, both cues are learned as moderately strong predictors of outcome E. Then, when learning I.PL → L, attention shifts away from cue I toward cue PL, and a strong link from PL to outcome L is acquired. Attention shifts away from cue I when learning I.PL → L because attention to cue I produces the wrong response, namely, outcome E.

A series of models that formalize attention shifting has been created by Kruschke (1996, 2001b, 2001c, 2006c) and extended to continuous stimuli by Kalish and Kruschke (2000; Kalish, 2001). The general framework of the models is displayed in Figure 1. Each cue has a multiplicative attentional gate, indicated in Figure 1 by triangles impinging upon the upward flow of cue activation. When attention on a cue is zero, then the cue activation is squelched. Each cue recruits some attention by its mere presence, but there can be competition for attention if there are multiple cues. A key aspect of the framework is that cue-outcome learning is actually indirect via two mappings: There is a learned mapping from cues to attentional allocation across the cues, and there is a learned mapping from attended cues to outcomes. These two distinct mappings are suggested by the curved arrows Figure 1.

The environment specifies which cues are present and which outcome is correct, but the environment does not specify how to allocate attention

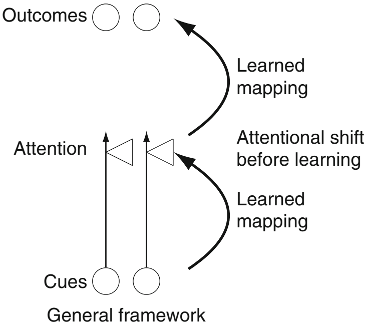


Figure 1 General framework for models of attentional shifting and learning.

across the cues. The models allocate attention to maximize the accuracy of the predicted outcome. In error driven, connectionist implementations of the framework (Kruschke, 1996, 2001b, 2001c), attention shifts away from cue I during the learning of I.PL → L because doing so reduces error. This attentional shift facilitates rapid acquisition of I.PL → L and reduces interference with the previously learned mapping I.PE → E. On a given trial, after attention has been shifted, then there is learning of the mappings from the presented cues to the attentional allocation, and from the attended cues to the outcomes. Attentional shifting and learning demonstrably improve performance on both early and late cases (Kruschke, 2003a).

In Bayesian implementations of the framework (Kruschke, 2006b, 2006c), attention shifts away from cue I during learning I.PL → L because doing so reduces inconsistency with the previously learned belief that cue I is associated with outcome E. After attention has been shifted on a given trial, the mappings are learned from the presented cues to the attentional allocation, and from the attended cues to the outcomes. In the Bayesian implementation, learning of a mapping entails shifting belief away from candidate mappings that are inconsistent with the training items, toward candidate mappings that are consistent with the training items (for an introduction to Bayesian associative learning, see Kruschke, 2008). Because of the attention shifting, this architecture for Bayesian learning robustly exhibits highlighting. An advantage of this Bayesian scheme over the connectionist models is that it can also exhibit phenomena known as "retrospective revaluation" (e.g., backward blocking, unovershadowing, etc.; see Kruschke, 2006c), which are very challenging to connectionist models but are naturally accommodated by Bayesian systems.

Because Bayesian learning of each mapping in Figure 1 is influenced only by its local information, the approach is called "locally" Bayesian learning. This learning scheme is different from standard Bayesian approaches in which both mappings are represented jointly in a global hypothesis space (e.g., Neal, 1996; Rumelhart, Durbin, Golden, & Chauvin, 1995). Globally Bayesian models do not exhibit highlighting because they are not sensitive to trial order. The locally Bayesian model is sensitive to trial order because the internal attentional targets, generated on a given trial to be least inconsistent with current beliefs, depend on previous learning.

Locally Bayesian learning is motivated generally by the idea that different levels of analysis may be Bayesian. Individual neurons might be Bayesian learners (e.g., Deneve, 2008), or committees of people might be Bayesian learners (cf. Akgün, Byrne, Lynn, & Keskin, 2007). Theories of mind typically posit numerous component processes, any of which could be Bayesian learners. For a mind to be globally Bayesian, it would have to keep track of all possible combinations of all possible states within and across components. This might be possible with clever algorithms, but it is more plausible to assume that each component keeps track of only its own possible states, and undergoes only locally Bayesian learning.

Another interesting behavior of locally Bayesian learning, when applied to highlighting, is that it learns faster than globally Bayesian learning (Kruschke, 2006c, p. 688). The retardation in the globally Bayesian system occurs because the global system must dilute its beliefs over a large number of possible joint hypotheses, and this uncertainty produces less decisive responses during learning.

In summary, the main point is that attention shifting is adaptive: Attention shifting accelerates acquisition of novel cases, and attention shifting preserves previous knowledge. The phenomenon of highlighting is a behavioral signature of this adaptive process. Highlighting is not an accidental deficiency in an otherwise well-tuned learning system. Highlighting is a direct consequence and sign of learning well.

2.3. Highlighting Is Robust, Pervasive, and Consequential

The claim, that highlighting is sign of learning well, is bolstered by the fact that it shows up in many places. It is not to be dismissed as a quirk, occurring by accident only in obscure and contrived conditions that have little relevance to most learning. The phenomenon is, in fact, robust and pervasive. And it has consequences predictable from attentional theory.

2.3.1. Robust

The effect persists under a variety of relative base rates, changes in base rates during training, different numbers of copies of the basic structure, different numbers of imperfect or perfect predictors, and so forth. For example, the designs of Medin and Edelson (1988) used three copies of the basic structure, but the designs of Medin and Bettger (1991) used four copies of the basic structure, and the designs of Kruschke (1996) used two copies. Robust highlighting was obtained in all the designs.

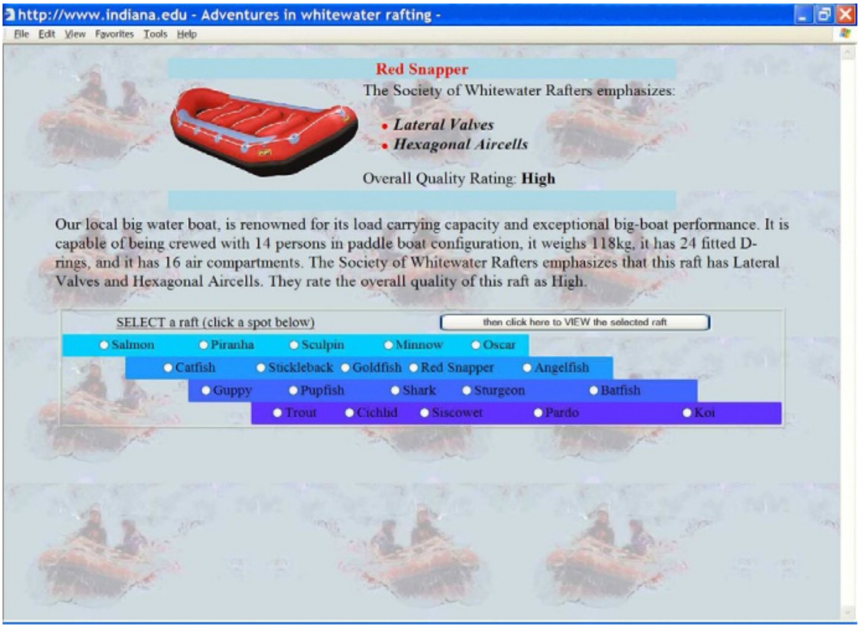
As another example, Medin and Edelson (1988, Experiment 2) reported a design in which one copy of the highlighting structure involved two shared predictors ($I_1.I_2.PE \rightarrow E$, $I_1.I_2.PL \rightarrow L$), a second copy had only one shared predictor but two perfect predictors for each outcome ($I.PE_1.PE_2 \rightarrow E$, $I.PL_1.PL_2 \rightarrow L$), and a third copy had no shared predictors ($PE_1.PE_2.PE_3 \rightarrow E$, $PL_1.PL_2.PL_3 \rightarrow L$). As anticipated by attentional theory, the magnitude of highlighting depended on the number of shared predictors. Kruschke (2001b, Experiment 1) also showed that a shared predictor was essential for producing highlighting, using a design with only two cues per outcome and only two copies of the basic structure.

Some researchers have used more extreme differential base rates than used in the original experiments by Medin and Edelson (1988). For example, Shanks (1992) and Juslin et al. (2001) used 7-to-1 base rates (instead of 3-to-1) and observed highlighting. The more extreme base-rate ratio could produce a stronger highlighting effect, presumably because it more strongly ensured that one outcome is well learned before the other outcome is learned.

Medin and Bettger (1991) explored changes in relative base rates during training. As long as one outcome had higher base rate than the other during initial training, thereby causing the high base-rate case to be learned before the low base-rate case, then the highlighting effect was observed. Subsequent designs have used only two copies of the basic structure, both changing base rates at the same time, and again found strong highlighting (Kruschke, 2001b; Kruschke, Kappenman, & Hetrick, 2005).

Across all these variations in design, the essential features seem to be that (i) one outcome is learned before the other outcome, (ii) the shared cue is associated with the early outcome, and (iii) the later learned outcome is well learned by test time. These design aspects are distilled into a ''canonical'' design described later in this chapter. An emphasis of the canonical design will be that highlighting does not depend on differential base rates overall; instead, the essential requirement is that one outcome is well learned before the other outcome is learned.

Various experiments have used different stimuli or cover stories or procedures. For example, in unpublished research conducted in 2001 by Kruschke with collaboration of an undergraduate honors student named Nancy Aleman, participants were instructed that they were to learn about the qualities of whitewater rafts. This knowledge could be used for decisions about which rafts to rent or purchase. Learners browsed 20 Web pages to learn about the rafts currently available on the market. Figure 2 shows an example of a Web page seen by the learners. Two features of the raft are given prominence in the display, along with the quality rating. The instance in Figure 2 features ''Lateral Valves'' and ''Hexagonal Aircells,'' with a ''High'' quality rating. These attributes might correspond to abstract cue I, cue PE, and outcome E, respectively. Additional text reiterates the features and quality in prose that was intended to imitate catalog sales descriptions. Notice that the pages did not require any explicit quality prediction for each case; learners merely read the information on the page. Participants selected whatever page they wanted to inspect next by selecting it from among the array of raft names at the bottom of the page. If a participant systematically selected rafts in reading order, that is, left to right and top to bottom, then they would encounter I.PE → E cases before I.PL → L cases. Most subjects did spontaneously select rafts in that order. Across all 20 pages, there were an equal number of E and L cases. After viewing all 20 pages, learners then viewed a few pages that purported to show prototypes of rafts that manufacturers were considering bringing to market. Participants predicted the quality of each raft based on the features of the raft. Results showed a strong highlighting effect in predicted quality: Rafts with the imperfectly predictive feature were given the earlier learned quality, and rafts with a combination of the two



http://www.indiana.edu - Adventures in whitewater rafting -
File Edit View Favorites Tools Help

Red Snapper

The Society of Whitewater Rafters emphasizes:

- *Lateral Valves*
- *Hexagonal Aircells*

Overall Quality Rating: **High**

Our local big water boat, is renowned for its load carrying capacity and exceptional big-boat performance. It is capable of being crewed with 14 persons in paddle boat configuration, it weighs 118kg, it has 24 fitted D-rings, and it has 16 air compartments. The Society of Whitewater Rafters emphasizes that this raft has Lateral Valves and Hexagonal Aircells. They rate the overall quality of this raft as High.

SELECT a raft (click a spot below)

○ Salmon ○ Piranha ○ Sculpin ○ Minnow ○ Oscar
○ Catfish ○ Stickleback ○ Goldfish ○ Red Snapper ○ Angelfish
○ Guppy ○ Pupfish ○ Shark ○ Sturgeon ○ Batfish
○ Trout ○ Cichlid ○ Siscowet ○ Pardo ○ Koi

Figure 2 Example of stimulus used for assaying highlighting in browsing a catalogue of whitewater rafts.

perfectly predictive features were given the later learned quality. These results show that overt predictive learning is not necessary for highlighting, nor is an austere "cues only" display.

Pictorial stimuli with joystick responses were used by Fagot, Kruschke, Dépy, and Vauclair (1998). Simple geometric figures, such as an oval or rectangle, were used to instantiate cues. The learner initiated a trial by using a joystick to move the cursor to the center of the screen. The cue figures would appear on the left or right of the screen. The learner made a response by moving the cursor to one of two colored squares that were positioned vertically above or below the start box. Learners were told merely to figure out which box to move to, in response to the various figures. The results again showed robust highlighting. Lamberts and Kent (2007, described in more detail below) also used pictorial stimuli and found robust highlighting. Although the stimuli might have been covertly named by the learners, these results show that textual stimuli are not necessary for highlighting to occur.

The effect has been found with socially relevant stimuli such as personality traits and group membership (Sherman et al., 2009; Wedell & Kruschke, 2001). In one design, Wedell and Kruschke (2001) trained people to predict a (fictitious) person's identity from his personality attributes. For example, the abstract case I.PE → E was instantiated as ''honest'' and ''conventional'' indicates ''Fred,'' and the abstract case I.PL → L was instantiated as ''honest'' and ''materialistic'' indicates ''Jack.'' The shared trait, ''honest,'' is known from previous norms to be a positive trait, while the distinctive traits, ''conventional'' and ''materialistic,'' are known to be equally negative traits. After learning to predict the person names from the traits, participants were asked to rate the likeability of each person. Presumably the rating of likeability is based on how strongly the traits have been associated with each person. If the traits were asymmetrically highlighted during learning, then the later learned person should be more strongly associated with the distinctive negative trait, and the earlier learned person should be more strongly associated with the shared positive trait. The actual likeability ratings confirmed this prediction, with the early learned person being rated more likable than the later learned person. Notice that ratings of likability use associations from outcomes (the person name) to cues (the traits), rather than from cues to outcomes, which suggests that the highlighting effect is caused by asymmetries in associations, not purely by biases at test.

The highlighting effect is modulated by cue salience, as anticipated by attentional theory. Continuing from the previous paragraph with the example of trait-name learning, Wedell and Kruschke (2001) found that if both the PE and PL traits were equally negative or equally positive, then a typical magnitude of highlighting was obtained. Previous literature strongly suggests that negative traits are more salient than positive traits. In other words, negative traits should attract attention more than positive traits. Consistent with this prediction, Wedell and Kruschke (2001) found that highlighting was magnified when the PL trait was negative while the PE trait was positive, and highlighting was diminished when the PL trait was positive while the PE trait was negative. Analogously, Bohil, Markman, and Maddox (2005) found that a highlighting-like effect could be generated if one distinctive cue were more salient than the other distinctive cue, even when the two outcomes were learned contemporaneously.

In a cued-recall paradigm, effects exactly analogous to highlighting were obtained by Dennis and Kruschke (1998). Learners saw two words such as ''digit'' and ''album'' at the top of a computer screen for 2 s, and were instructed to covertly anticipate the word, such as ''shark,'' that would appear after a pause at the bottom of the screen. Learners simply watched a sequence of such trials before a test phase in which words appeared at the top of the screen and the anticipated word had to be typed on the computer keyboard. This procedure is interestingly different from the standard predictive-learning paradigm. First, there is no explicit feedback regarding the correctness of the covertly anticipated response during learning. Second, there is no cover story relating the cues to the outcomes, and no causal relationship such as that between diseases and symptoms (used in previous research by Kruschke, 1996; Medin & Bettger, 1991; Medin & Edelson, 1988). Third, and perhaps most importantly, the response in the test phase is not limited to the words seen as outcomes during learning, because participants could type in any word at all, including the cue words or any other word that came to mind. Despite these differences, the test results were remarkably consistent with the results from previous predictive-learning experiments, revealing robust highlighting. Thus, forced choice at test is not required for highlighting to occur.

The highlighting effect is also robust under time pressure and a dual task during test. After training with pictorial stimuli in a design using fixed 3-to-1 base rates, Lamberts and Kent (2007) tested participants in four different conditions. One test condition allowed the usual unpressured response. A dual-task condition had subjects simultaneously counting quickly backward in multiples of three during the test responses. Two other conditions demanded responses be given within 500 or 300 ms. The signature torsion of highlighting was clearly obtained in all four test conditions, merely somewhat attenuated in the speeded conditions. Lamberts and Kent (2007) argued that the robustness of highlighting under time pressure and a dual task made it unlikely that highlighting can be fully explained by inferential rules executed at time of test, because rule application should be disrupted by those additional cognitive demands.

2.3.2. Highlighting Is Correlated with Blocking and Gaze

Learning in other cue-outcome designs should also be affected by attentional shifting. One such design, known as "blocking" (Kamin, 1968; Shanks, 1985), trains people in an early phase with cases of A → 1, then in a later phase with cases of A.B → 1. In other words, the later phase introduces a perfectly predictive cue which is redundant with an already learned cue. As a comparison, the later phase also has cases of C.D → 2, without any earlier training of cues C or D. In test, the redundant cue is put in conflict with a comparison cue: B.D → ? People prefer outcome 2, which suggests that learning about cue B was attenuated, or "blocked," because of the previous learning about cue A.

One explanation of blocking is that it is caused, at least in part, by learned *in*attention to the blocked cue (e.g., Kruschke & Blair, 2000; Mackintosh & Turner, 1971). The idea is that during learning of A.B → 1, cue B distracts attention away from the already predictive cue A. This distraction causes diminished accuracy. To alleviate the error, attention is redirected back to cue A, away from cue B. Thus, people learn to suppress attention to cue B.

Because the same attentional shifting process is supposed to be at work in both blocking and highlighting (but yielding complementary effects), the effects should be correlated. In other words, a person who has especially strong attentional shifting should show relatively strong blocking and highlighting, but a person who has relatively small attentional shifting should show a lesser degree of blocking and highlighting.

This predicted correlation was verified by Kruschke et al. (2005). People were trained on both blocking and highlighting designs, and magnitudes of blocking and highlighting were estimated for each person from their choice preferences at test. Across people, there was a significant positive correlation between blocking and highlighting.

The correlation of blocking and highlighting is another challenge to models of the phenomena. The attentional shifting and learning model of Kruschke (2001c) was shown to accommodate the correlation (Kruschke et al., 2005). Specifically, when the attentional parameters of the model are varied, to mimic individual differences in attentional shifting, the model naturally predicts covarying magnitudes of both blocking and highlighting. Importantly, variations in other parameters, such as associative learning rates or choice decisiveness, do not account for the covariation. Other models have difficulty addressing this correlation.

Attentional theory asserts that covert attention is directed at the highlighted cue. If overt eye gaze reflects covert attention, then gaze should dwell for longer duration on highlighted cues and for shorter duration on blocked cues. This prediction was confirmed by Kruschke et al. (2005). Figure 3 shows an example of a gaze trajectory on a single trial, where the clusters of dots indicate places where the eyes fixated. Moreover, individual measures of differential gaze durations correlated with differential choice preferences. In other words, people who showed stronger blocking and highlighting in their choice preferences tended to show greater differences in gaze durations for blocked or highlighted cues. The correlations between choice preferences and gaze differences, and between blocking and highlighting, are shown in the lower part of Figure 3. Although the correlation of choice and gaze is not a necessary prediction of attentional theory, because it is based on the additional assumption that overt gaze follows covert attention, the correlation of blocking and highlighting is a fairly firm prediction, qualified only by the independent variation induced by other influences.

2.3.3. Learning *After* Highlighting

Attentional theory posits that learners rapidly reallocate attention across cues when the default allocation causes inaccurate prediction. This idea is annotated in Figure 1 as ‘‘Attentional shift before learning.’’ A further premise is that people learn these reallocations of attention, so that on subsequent repetitions of the same cues, attention can be more appropriately apportioned and yield more accurate responses. This idea is annotated in Figure 1 as ‘‘Learned mapping’’ from cues to attention.

If attentional allocations are learned in highlighting and blocking, then the learned allocations should take time to overcome if the cue-outcome

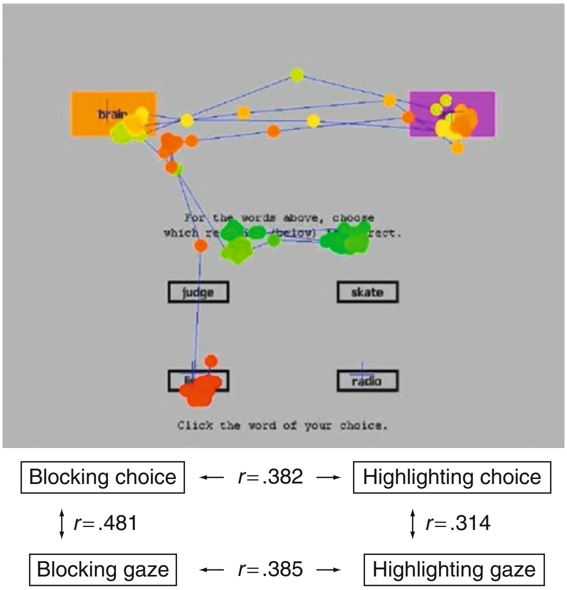


Figure 3 Top: Example of a gaze trajectory in experiment by Kruschke, Kappenman, and Hetrick (2005). Each dot indicates gaze location as sampled at 1/60 s, with lines connecting consecutive locations. The trajectory begins with the dots near the center over the instructions, moves up to the dots over the cue words at the top of the display, and then moves down to the dots over the response button that was clicked with the screen cursor. Bottom: Correlations, across individuals, of magnitude of blocking or highlighting, assayed by choice or gaze preference.

mapping changes in subsequent training. Specifically, if people have learned to ignore a blocked cue, then learning a new association from the blocked cue should be relatively difficult in subsequent training. This prediction was confirmed in different experiments reported Kruschke and Blair (2000) and Kruschke (2005).

Moreover, if people have learned to attentionally highlight cue PL in the highlighting design, then it should be difficult to ignore the highlighted cue if it becomes irrelevant in subsequent training. This prediction was also confirmed by Kruschke (2005). The experiment trained people in a typical highlighting design and then continued training with new outcomes. For both of two groups of learners, the new outcomes were perfectly indicated by the cues that had been the imperfect I cues during highlighting. For one group of learners, the I cues were accompanied by the PE cues, which were randomly paired with the novel outcomes, but for the other group of learners, the I cues were accompanied by the PL cues, again randomly paired with the novel outcomes. If there was stronger learned attention to the PL cues than to the PE cues, then learning to ignore the newly irrelevant

PL cues should be more difficult than learning to ignore the newly irrelevant PE cues. This difference in learning difficulty was in fact observed in the data.

2.3.4. Highlighting with Continuous Cues and Outcomes

The previous sections discussed experiments in which the cues were present/absent features, such as words or geometric figures. In the real world, cues can instead have continuous magnitudes, not just binary discrete states. In recognition of these alternative cue instantiations, some researchers have sought effects analogous to highlighting with continuous cues, and even continuous outcomes.

Kalish and Kruschke (2000) considered continuous-dimension stimuli with categorical outcomes. The cue-outcome structure was not intended as a direct analogue of highlighting, but was intended to examine shifts of attention among values within a dimension. The hypothesis was that early learned stimuli would be encoded in terms of their typical, average values. Later learned stimuli, of different categories, would be encoded by more extreme stimulus values, because those extreme values would help discriminate the new category from the previously learned one. A model was developed that incorporated attentional shifts across dimensions, but also attentional shifts across values within dimensions. The model was able to mimic some subtle effects in people's choice preferences across the continuum of stimulus values, but those subtleties could not be captured when the attention shifting in the model was disabled.

Kalish (2001) considered a design directly analogous to highlighting, in which the present/absent values of the standard experiment were instantiated as two different values on a continuum; such as short and tall heights of a vertical bar. In different experiments, the heights had different amounts of random noise added. When the random variation did not cause categories to overlap in stimulus space (i.e., when there was a deterministic mapping from stimuli to outcomes), highlighting was obtained. Kalish modeled the results with an extension of the attention-shifting model presented by Kalish and Kruschke (2000).

When both the cues and outcomes are continuously valued, the design falls into the realm of "function" learning. This appellation in the literature is apparently based on analogy to high-school mathematics, wherein mappings from continuous x to continuous y, such as lines and higher order polynomials, are the prototypical functions. If function learning and discrete-outcome learning are related, it is plausible that attentional shifts should occur during function learning. Suppose, for example, that cues I, PE, and PL are instantiated as continuous valued dimensions, such as body temperature, grade point average, and hair length. People are trained early with cases in which small values of I and PE lead to small values of the outcome, and other cases in which large values of I and PE lead to large values of the outcome. Thus, the I and PE values covary across trials and indicate correspondingly covarying outcome values. Later in training, people are trained with cases in which values of I and PL covary, but the outcome is *negatively* correlated with the cue values. In other words, the function relating I.PE cases to the outcome is positive linear, but the function relating I.PL cases to the outcome is negative linear. If attentional highlighting occurs, then, when tested with cue I by itself, people should prefer to give positive linear responses, but, when tested with cues PE.PL, people should prefer to give negative linear responses. This pattern of responding was in fact observed in an experiment reported by Kruschke (2001a). The analogous result for blocking was also found.

In summary, highlighting does not seem to be limited to discrete cues and outcomes. Although there has been relatively little investigation of highlighting with continuous cues, it seems advisable that theories of highlighting should be extendible, in principle at least, to continuous cues and outcomes.

2.3.5. Possible Sightings Afield

A variety of phenomena in other domains have been addressed by attentional theories much like the one that accounts for highlighting. These phenomena tend to share two main qualities. First, learning of new items can be fast. This rapidity can be explained, at least in part, by the ability of attention to focus on distinctive features or representations that reduce interference with previously learned items. Second, learned knowledge can be distorted relative to the actual stimulus statistics. This distortion also can be explained, at least in part, as the consequence of selective attention that is differently tuned for different items at different points of learning. These ideas have been applied to aspects of *language acquisition* (Colunga & Smith, 2008; Ellis, 2006; Goldberg, Casenhiser, & Sethuraman, 2005; Parish-Morris, Hennon, Hirsh-Pasek, Golinkoff, & Tager-Flusberg, 2007; Regier, 2005; Smith & Yu, 2008; Yoshida & Hanania, 2007), *consumer learning* (Cunha, Janiszewski, & Laran, 2008; Cunha & Laran, 2009; Kruschke, 2006a; Pieters, Warlop, & Wedel, 2002; van Osselaer & Janiszewski, 2001), *context cues* in learning (Nelson & Callejas-Aguilera, 2007; Rosas & Callejas-Aguilera, 2006; Rosas, Callejas-Aguilera, Ramos-Álvarez, & Abad, 2006), and learning in *social cognition* (e.g., Cramer et al., 2002; Hayes, Foster, & Gadd, 2003; Sherman et al., 2009; Wedell & Kruschke, 2001), among others. There is not space here to discuss all these connections to the literature, but it is hoped that these pointers are suggestive of the potential scope of highlighting in learning.

2.4. Interim Summary and Goal of Remainder

In summary, the highlighting effect has been found with a variety of stimuli, cover stories, stimulus frequencies, and cue combinations. It is correlated with blocking, and it has predictable consequences for subsequent learning. Highlighting is not merely a stubborn deficiency of otherwise rational learning; rather, highlighting is adaptive because it reduces interference with previous knowledge and accelerates acquisition of new knowledge. The style of attentional theory that explains highlighting has been applied in a variety of domains. Thus, the highlighting phenomenon is among the catalog of major phenomena that learning theories need to address. Other researchers agree: "Because [highlighting] is so problematic, we will argue that the effect goes to the heart of several important issues in human learning and decision making" (Johansen, Fouquet, & Shanks, 2007, p. 1366).

The primary goal for the remainder of this chapter is to present new results from a "canonical" highlighting experiment that may serve as a test bed for models of learning. In a canonical design, the overall frequencies of the early and late cases are equal. In other words, there is no overall difference in base rates. The canonical design also has an initial phase in which the early cases are trained without any interspersed late cases, thereby guaranteeing that the early cases are actually learned before the late cases. One purpose of this canonical design is to demonstrate unambiguously that the highlighting effect does not depend on overall differences in bases rates; that is, the highlighting effect is not (only) an inverse base-rate effect, because there are no overall base rate differences to invert. Another purpose of the canonical design is provide concrete data that challenge models that rely on differential base rates to account for the highlighting effect. Such models include some recent Bayesian approaches, including the Rational model (Anderson, 1990, 1991) and the Kalman filter model (Dayan et al., 2000; Kruschke, 2008). Finally, the data are analyzed using Bayesian methods, unlike all previous reports in the literature. The hierarchical Bayesian analysis allows for individual differences, and it provides a complete posterior distribution of credible response preferences.

3. Experiment: A "Canonical" Design with Equal Base Rates

A framework for a "canonical" design for highlighting was suggested by Kruschke (2006c, Table 1, p. 686). The design is guided by three motivations. First, some exposures to I.PE → E should occur initially, so that it is definitely learned first. Second, the total number of cases of I.PE → E should equal the total number of cases of I.PL → L. Third, aside from the

Table 1 A Canonical Highlighting Design.

Phase	# blocks	Item × Frequency	
First	N_1	I.PE → E × 2	
Second	N_2	I.PE → E × 3	I.PL → L × 1
Third	$N_3 = N_2 + N_1$	I.PE → E × 1	I.PL → L × 3
Test	2	I.PE → E × 2 I → ? × 1	I.PL → L × 2 PE.PL → ? × 1, etc.

Note: An item is shown in the format, Cues → Correct Response × frequency per block. The actual experiment has two copies of the structure shown here; for example, the first phase involves I1.PE1 → E1 × 2 and I2.PE2 → E2 × 2.

initial training, the relative base rates should never be too extreme, because people should be learning about the cases in relation to each other. The 3-to-1 base rates established by Medin and Edelson (1988) and by Medin and Bettger (1991) were used as a guideline.

Table 1 shows a canonical highlighting design. It has three phases of training. The first phase presents only cases of I.PE → E, to ensure that at least some learning of the early cases does happen before the later cases. A block of the first phase involves two repetitions of I.PE → E, and there are N_1 blocks. The second phase introduces the cases of I.PL → L, but at only one third the frequency of the early cases. There are N_2 blocks of the second phase. The third phase reverses the base rates, emphasizing the later learned cases. The third phase has N_3 blocks. Within all blocks, the trials are permuted randomly. The blocks progress seamlessly without any marker between blocks.

Notice in the table that when $N_3 = N_2 + N_1$, the total number of I.PE → E trials is 3 N_1 + 4 N_2, which equals the total number of I.PL → L trials. This equality of base rates distinguishes highlighting from the inverse base rate effect reported by Medin and Edelson (1988), which used only the second phase of Table 1, i.e., $N_1 = 0$ and $N_3 = 0$. One possible infelicity of the canonical design proposed here is that training ends with one outcome occurring more often than the other, and this short-term imbalance in favor of the later trained outcome may carry over into the test items. To solve this problem, a fourth training phase could be appended (still before the test phase) in which the two cases are interspersed with equal frequency. Such a candidate fourth phase was not used here for two reasons. First, the test phase includes continued interspersed training with equal base rates, as shown Table 1, albeit with only a modest number of repetitions. Second, Medin and Bettger (1991, Experiment 2) showed that training with equal base rates after an initial phase with 3-to-1 base rates still produced the signature torsion of highlighting.

Medin and Bettger (1991) reported experiments in which one subset of the cue-outcome pairs had balanced frequencies, corresponding with $N_1 = 0$, $N_2 = 6$, and $N_3 = 6$. The canonical design instead has $N_1 > 0$ to assure that the early items really are learned before the later items. The designs used by Medin and Bettger (1991) also interleaved learning of balanced structures with imbalanced structures, leaving open the possibility that learning of an imbalanced structure influenced the learning of a balanced structure.

The canonical design does not require the number of blocks to be fixed in advance. Instead, training can continue in the first and second phases until an accuracy criterion is reached. For example, training in phase 1 could continue until accuracy achieves 11/12 in a window of three consecutive blocks (as was done by Kruschke, Kappenman, & Hetrick, 2005), and training in phase 2 could continue until accuracy on the later items achieves 5/6 in a window of three consecutive blocks. With the number of blocks in the first two phases, N_1 and N_2, established by the subject's achievement of criterial accuracy, the number of blocks for the third phase is set as $N_3 = N_1 + N_2$, thereby achieving overall balance of base rates while also assuring early learning of one case and high accuracy overall. Moreover, by using an accuracy criterion, the framework of the design can be applied to different stimuli, situations, and subjects, in which or for whom learning may be more or less difficult. The experiment reported below, however, used a fixed number of blocks, merely to maintain consistency with previously reported experiments.

The equality of base rates in the canonical design emphasizes that highlighting is an order-of-learning effect, not a base rate effect. It is only by virtue of the fact that the I.PE cases are learned before the I.PL cases that asymmetric responding occurs at all. If the I.PE and I.PL cases were intermixed equally throughout training, they would be structurally equivalent and no such highlighting effect could be meaningfully assayed (except for idiosyncratic differences in acquisition order by individual subjects).

3.1. Method

3.1.1. Design

The canonical design of Table 1 was used with $N_1 = 10$ and $N_2 = 5$. There were two copies of the basic design intermixed, so that the first phase involved more than one correct response. Hence there were a total of six cues and four outcomes. This yielded a total of 200 training trials. Across the 200 training trials, there were 50 trials of each of the I1.PE1 → E1, I1.PL1 → L1, I2.PE2 → E2, and I2.PL2 → L2 items. The order of items was randomly permuted within each block.

The testing phase continued seamlessly after the training phase. Each testing block contained two trials of each of the four training items with feedback, as indicated in Table 1. This continued training equalized the short-term base rates during test, served as a reminder of the correct outcomes in the midst of the test trials, and simultaneously assessed accuracy on the training items. Each testing block also contained the 11 other test types shown in Table 2. Each of the 11 test types was probed once per block for each copy of the cue structure. Therefore each test block contained 30 trials, which were randomly permuted within blocks. There were two test blocks. The totality of the experiment comprised 260 trials, and took approximately 15 min for a participant to complete.

3.1.2. Stimuli

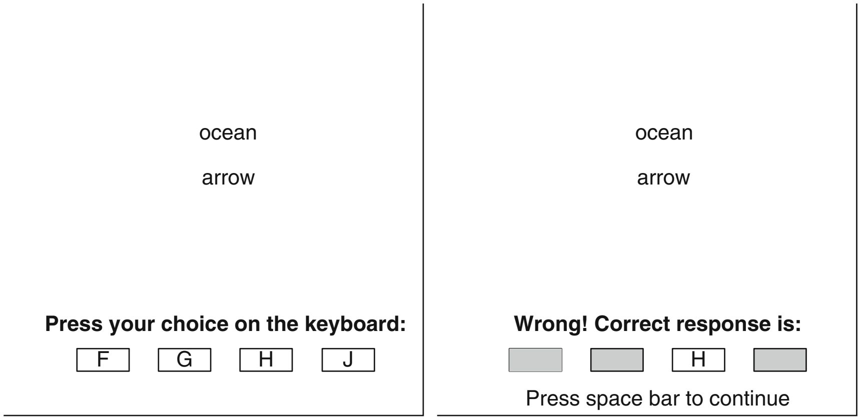
The six cues words were either the set, "snake," "robin," "whale," "puppy," "skunk," and "trout," or else the set, "child," "mouse," "ocean," "tulip," "piano," and "arrow." These words were selected because they are highly concrete and imagable, they all have five letters, they all have different initial letters (within a set) that are also different from the letters of the response keys, and there are no striking semantic relationships between words (within a set). The set used for a participant was selected randomly. The assignment of the six words to the six abstract cue types was randomly permuted for each participant.

The response keys were F, G, H, and J. These are the four central keys on a standard keyboard. Figure 4 shows examples of the stimuli as displayed

Table 2 Response Percentages for Each Probe in the Test Phase of the Canonical Highlighting Design.

	Response			
Cues	E	L	Eo	Lo
I.PE	91.8	5.9	1.0	1.4
I.PL	3.9	93.9	1.4	0.8
I	63.7	26.2	6.2	3.9
PE.PL	35.2	57.8	3.5	3.5
PE.PLo	29.3	5.9	5.1	59.8
PE	85.9	5.1	5.1	3.9
PL	3.9	87.5	5.5	3.1
I.PE.PL	43.8	45.7	3.5	7.0
I.PEo.PL	13.3	62.5	17.6	6.6
I.PE.PLo	48.4	7.8	5.5	38.3
I.PEo.PLo	9.8	16.0	27.7	46.5
I.PEo	21.9	19.9	51.6	6.6
I.PLo	11.7	16.8	3.1	68.4

Note: The first two rows are based on 8 trials/subject, and the remaining rows are based on 4 trials/subject, with 64 subjects.



ocean

arrow

Press your choice on the keyboard:

ocean

arrow

Wrong! Correct response is:

Press space bar to continue

Figure 4 Left: Computer display for cues with response prompt. Right: Computer display for cues with corrective feedback. The actual displays had gray backgrounds, rendered here as white.

on the computer screen. The assignment of keys to abstract labels (E1, L1, E2, and L2) was randomly permuted for each participant.

Figure 4 shows examples of the stimuli and feedback as actually presented on the computer screen. The position of the cue words was randomly permuted from trial to trial. For example, on some I.PL → L trials, cue I appeared above cue PL, but on other trials, cue PL appeared above cue I.

3.1.3. Instructions

The instructions to the participant provided no causal cover story. For example, there was no mention of symptoms and diseases that several previous studies used (e.g., Kruschke, 1996; Medin & Bettger, 1991; Medin & Edelson, 1988). The instructions were neutral, saying only the following:

In this experiment you will see some common words on the computer screen. Your job is to learn which words indicate which keys to press. You can press "F," "G," "H," or "J." When the words are presented, you make a guess by pressing one of the keys. Please locate the F, G, H, and J keys on the keyboard now—they are in the middle of the keyboard. After you make your choice in response to the words, the correct answer will be displayed. At first you will just be guessing, but after several repetitions you can learn which words indicate which keys. The correct keys for the words never change, so you can achieve perfect accuracy if you try. At some times during the experiment, new words may be introduced. Just learn these new words as accurately as you can.

3.1.4. 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 72 participants.

3.2. Results

3.2.1. Learning Criterion

The results on the generalization probes are only of interest if the participants accurately learned the training items. If chance performance is considered to be 1/4 correct, because there were four response options, then significantly above chance requires 6 out of 8 correct (two-tailed, $p < 0.05$).[2] Therefore, if a participant showed fewer than 6 out of 8 correct responses on either I.PE or I.PL trials in the test phase, he or she was excluded from further analysis. The learning criterion eliminated only 8 of 72 participants (i.e., 11%), leaving $N = 64$.

3.2.2. Choice Data

Table 2 shows the average percentage of choices of each response category, for all the different test items. Each cue had outcomes with which it was associated during training, and *other* outcomes with which it was not associated. For example, during training, there occurred cases of I1.PL1 → L1 and I2.PL2 → L2. In test, there were probes involving combinations of cues from different sets, such as I1.PL2 and I2.PL1. Because of the structural symmetry in the design, these cases were collapsed and denoted I.PLo, with the lowercase "o" indicating the *other* copy of the cues. Responses corresponding with the other cue were also marked with an affixed lowercase "o." For example, if the probe is I1.PL2 and the response is L2, the probe is tabulated as a case of I.PLo with response Lo. If the response to I1.PL2 is instead E1, it is tabulated as a case of response E.

First, notice that accuracy for the training items was very high in the test phase. (Recall the learning criterion demanded 6 out of 8, i.e., 75% correct, on both items.) The first two rows of Table 2 indicate that performance on the training items was in the low-nineties percent correct.

[2] The learning criterion can be motivated from a Bayesian perspective instead of from null hypothesis significance testing. Suppose the prior belief regarding the underlying probability correct on training items has a mean of 1/4, that is, guessing, but has a large uncertainty, expressed as a beta(1, 3) distribution. When 6 of 8 test trials are correct, the resulting posterior beta(6 + 1, 2 + 3) distribution has a 95% HPD interval (from 0.318 to 0.841) that excludes the chance value 0.25. But when only 5 of 8 test trials are correct, the posterior beta(5 + 1, 3 + 3) distribution has a 95% HPD interval (from 0.234 to 0.766) that includes the chance value 0.25. The same conclusion is reached if the prior is beta(2, 6), instead of beta(1, 3), which expresses somewhat higher prior certainty that the learner is merely guessing.

The results show a robust highlighting effect. For the imperfect cue I (third row of Table 2), there was a strong preference for response E over response L, with people selecting E more than twice as often as L (63.7 vs 26.2%). Statistical analyses are provided in Section 3.2.3. On the other hand, for the conflicting-cue case of PE.PL (fourth row of Table 2), there was a robust preference for response L over response E. Thus, the trademark "torsion" of highlighting is strongly displayed in this canonical design.

The dominance of PL over PE is also revealed by several other test probes. Probe PE.PLo (fifth row of Table 2) shows that response Lo is preferred over response E. Probe I.PEo.PLo (third from bottom row of Table 2) shows that Lo is preferred over Eo. And comparing I.PEo with I.PLo (bottom two rows of Table 2) shows that PLo dominates I more than PEo dominates I.

The remaining probes are included primarily to fill out all possible cue combinations (with up to three cues), for thoroughness and as additional constraints for future model fitting.

3.2.3. Bayesian Statistical Analysis

The data were analyzed using Bayesian methods. The appendix provides a few general reasons to prefer Bayesian methods over traditional null hypothesis significance testing. For the specific application here, traditional chi-square tests, which have been used in previous reports, are problematic because it is unclear how to combine data across subjects. Previous analyses have made the implausible assumption that all subjects are equally representative of a mutual übersubject, without any allowance for individual differences. Moreover, the traditional chi-square analyses merely test a null hypothesis of equal responding, without providing an estimate of what range of response biases are tenable, given the data. Both of these problems are addressed by the Bayesian analysis.

In the Bayesian analysis, a descriptive model of the data is defined, and the parameter values of the model are estimated. The Bayesian analysis yields a degree of belief in all possible parameter values, not merely a single best-fitting parameter value. 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.

The left side of Figure 5 suggests the basic structure of the model, informally. At the bottom left of the diagram, each individual's observed response frequencies are a random sample from that individual's underlying response propensity. The downward arrow in Figure 5 represents the generation of responses based on underlying propensities. Moving up a level, each individual's underlying response propensity is considered to be a random draw from some overall distribution of response propensities,

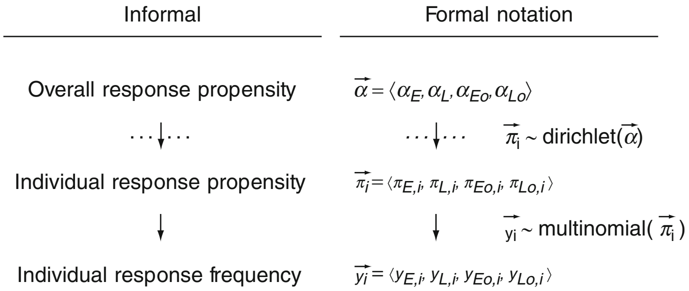

Informal
Formal notation
Overall response propensity
$\vec{\alpha} = \langle \alpha_E, \alpha_L, \alpha_{Eo}, \alpha_{Lo} \rangle$
$\vec{\pi}_i \sim$ dirichlet($\vec{\alpha}$)
Individual response propensity
$\vec{\pi}_i = \langle \pi_{E,i}, \pi_{L,i}, \pi_{Eo,i}, \pi_{Lo,i} \rangle$
$\vec{y}_i \sim$ multinomial($\vec{\pi}_i$)
Individual response frequency
$\vec{y}_i = \langle y_{E,i}, y_{L,i}, y_{Eo,i}, y_{Lo,i} \rangle$

Figure 5 **Hierarchical model analyzing each test item. Not indicated in the diagram is the hyperprior on α, which was $\alpha_r \sim$ gamma (0.25, 0.0025) r. See main text for discussion.**

governed by the cue combination. In other words, the cue combination at test evokes some overall response propensity. That overall response propensity has somewhat different manifestations in different individuals. The variation across individuals, due to distinct draws from the overarching propensity, is represented by the downward arrow with ellipses on either side.

The informal structure on the left side of Figure 5 is given formal precision on the right side of the figure. Notation will be explained from the bottom up. Recall that in the test phase, each cue combination was presented to the learner several times. For example, the test cue I was presented to each learner four times, and the four responses might comprise 2 E's, 1 L, 1 Eo, and 0 Lo's. These response frequencies, for the *i*th individual, are denoted by $\vec{y}_i = \langle y_{\mathrm{E},i}, y_{\mathrm{L},i}, y_{\mathrm{Eo},i}, y_{\mathrm{Lo},i} \rangle$, on the lowest row in Figure 5.

The particular response frequencies are modeled as a random sample from the individual's underlying response propensities, denoted $\vec{\pi}_i = \langle \pi_{\mathrm{E},i}, \pi_{\mathrm{L},i}, \pi_{\mathrm{Eo},i}, \pi_{\mathrm{Lo},i} \rangle$ in Figure 5. Mathematically, a random sample of categorical responses is generated by a multinomial distribution with underlying probabilities $\pi_{\mathrm{E},i}$, $\pi_{\mathrm{L},i}$, $\pi_{\mathrm{Eo},i}$, and $\pi_{\mathrm{Lo},i}$ (which sum to 1), and this sampling is denoted $\vec{y}_i \sim \text{multinomial}(\vec{\pi}_i)$.

An individual's response propensities are assumed to be a random representative of the overall response propensity induced by the test item. The overall response propensity for a test item is denoted $\vec{\alpha} = \langle \alpha_{\mathrm{E}}, \alpha_{\mathrm{L}}, \alpha_{\mathrm{Eo}}, \alpha_{\mathrm{Lo}} \rangle$ in the top row of Figure 5. Mathematically, a random sample of response probabilities is generated from a Dirichlet distribution that has parameters α_{E}, α_{L}, α_{Eo}, and α_{Lo}.

In summary, this hierarchical model allows individual differences to be captured by participant-level multinomial probabilities, mutually constrained by being drawn from the same higher level Dirichlet distribution which describes across-subject response tendencies for the cues.

The primary goal of the analysis is to generate a posterior estimate of overall response propensities $\vec{\alpha}$ for each probe item. For example, suppose that the estimated posterior distribution on the $\vec{\alpha}$ parameters for test cue I has a typical value of $\langle \alpha_{\mathrm{E}}, \alpha_{\mathrm{L}}, \alpha_{\mathrm{Eo}}, \alpha_{\mathrm{Lo}} \rangle = \langle 300, 500, 100, 100 \rangle$. This implies that typical individual-level response probabilities will be near $\pi_{\mathrm{E},i} = \alpha_{\mathrm{E}}/\Sigma\alpha_k = 0.30$, $\pi_{L,i} = \alpha_L/\Sigma\alpha_k = 0.50$, $\pi_{\mathrm{Eo},i} = \alpha_{\mathrm{Eo}}/\Sigma\alpha_k = 0.10$, and $\pi_{\mathrm{Lo},i} = \alpha_{\mathrm{Lo}}/\Sigma\alpha_k = 0.10$. The posterior distribution on the α parameters yields the explicit posterior probability that, for example, the response tendency for L is greater than the response tendency for E. This will be explained in more detail below, with the actual posterior distributions of specific probe items.

The posterior on the α parameters is also indicative of the across-subject consistency of responses. If all subjects give the same distribution of responses to a cue, then the α estimates are high, because high α's yield little variation among individual π_i values. But if subjects give responses that vary a lot from one person to another, then the posterior α values are low, to allow for variation among individual π_i values.

The prior distribution on the α parameters was set to be vague and equal for all components. Specifically, the prior on each α was a gamma density with mean 100 and standard deviation 200 (yielding gamma parameters of shape = 0.25 and rate = 0.0025). Because very small values of α caused trouble for the Dirichlet sampling, the gamma distributions were censored at 0.3. Small changes in the arbitrary censoring value yielded trivial changes in the posterior. This vague and unbiased prior was selected in an attempt to be unobjectionable to a general skeptical audience. (If the prior were instead informed by previously published results from related designs, such as those of Medin and Bettger (1991), then the posterior distributions reported below would be even stronger.)

The posterior distribution was determined by Markov chain Monte Carlo (MCMC) approximation. The simulations used the software BRugs, which is an R-language interface to OpenBUGS, which in turn is based on WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Five parallel MCMC chains were simulated, using a burn-in of 50,000 steps and thinning of 1000 steps. This extensive burn-in and thinning produced well-mixed chains with small auto-correlation, so the posterior sample is very trustworthy. From each of the five chains, 1000 steps were retained to represent the posterior, yielding 5000 representative parameter values.

Figure 6 shows results for selected cues. The upper panel shows results from cue I. The Bayesian analysis yielded 5000 representative points $\langle \alpha_{\mathrm{E}}, \alpha_{\mathrm{L}}, \alpha_{\mathrm{Eo}}, \alpha_{\mathrm{Lo}} \rangle$ in the posterior distribution. Each of those points indicates a credible combination of α values, given the data. At any of the 5000 points, the estimated overall probability that participants give a response of E is $p(E\,|\,I) = \alpha_{\mathrm{E}}/(\alpha_{\mathrm{E}} + \alpha_{\mathrm{L}} + \alpha_{\mathrm{Eo}} + \alpha_{\mathrm{Lo}})$. The preference of response E compared to response L is, therefore, $p(E\,|\,I) - p(L\,|\,I) = (\alpha_{\mathrm{E}} - \alpha_{\mathrm{L}})/(\alpha_{\mathrm{E}} + \alpha_{\mathrm{L}} + \alpha_{\mathrm{Eo}} + \alpha_{\mathrm{Lo}})$.

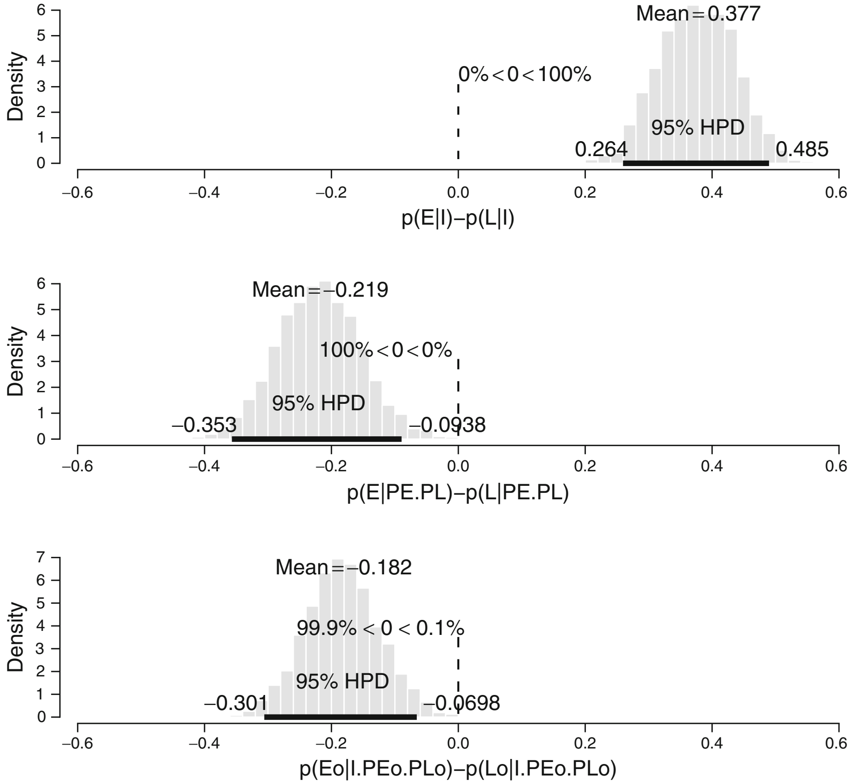


Figure 6 Posterior distributions of response biases for selected cues.

This preference is computed at each of the 5000 points in the posterior, and a histogram of those credible preferences is shown in the top panel of Figure 6. It shows that the mean estimate of $p(E|I) - p(L|I)$ is 0.377, the 95% highest posterior density[3] (HPD) region falls well above zero (ranging from 0.264 to 0.485), and 100% of the believable values are greater than zero. In other words, our posterior beliefs about the response bias for cue I are very firmly that E is preferred over L.

Analogously, the middle and lower panels of Figure 6 indicate that our posterior beliefs are very firmly that response L is preferred to response E for cues PE.PL, and Lo is preferred to Eo for cues I.PEo.PLo. The posterior

[3] By definition, all the points of the 95% HPD region have higher believability than points outside the region, and the region covers 95% of the believable values.

distributions reveal in detail just how believable are various magnitudes of preference. In particular, for cues I.PEo.PLo, the mean of believable values for $p(Eo\,|\,I.PEo.PLo) - p(Lo\,|\,I.PEo.PLo)$ is −0.182, the 95% HPD is well below zero (ranging from −0.301 to −0.0698), and 99.9% of the values are below 0. This bias for cues I.PEo.PLo is especially challenging for the eliminative inference model (Juslin et al., 2001; Kruschke, 2001b).

Another test cases of interest, not shown in Figure 6, is PE.PLo. The posterior for $p(E\,|\,PE.PLo) - p(Lo\,|\,PE.PLo)$ has mean −0.304, 95% HPD ranging from −0.433 to −0.171, and has 100% of its values less than 0. Finally, it is interesting to compare $p(Eo\,|\,I.Eo)$ with $p(Lo\,|\,I.PLo)$. The posterior for $p(Eo\,|\,I.Eo) - p(Lo\,|\,I.Lo)$ has mean −0.161, 95% HPD ranging from −0.253 to −0.067, and has 99.8% of its values less than 0.

The posterior distributions of α parameters are available on the author's Web site. The distributions are useful for two purposes. First, interested readers can examine the believable response propensities for all the test cases. Second, researchers who want to repeat the experiment can use the distribution as a prior for their own analyses.

In summary, the Bayesian analysis indicates that the signature "torsion" of highlighting is highly credible for these data from a canonical highlighting design. The Bayesian analysis avoided the questionable assumption of traditional chi-square analyses, that all individuals have the same response propensities. And unlike chi-square analyses that only indicate whether or not a null hypothesis can be rejected, the Bayesian analysis explicitly reveals the believabilities of various degrees of response preference.

3.3. Implications and Discussion

The main point of the results is that the classic highlighting effect occurs robustly even in a "canonical," equalized base-rates design. In other words, the highlighting effect is not properly called an inverse base-rate effect, because there are no differential base rates to invert. Indeed, as argued by Kruschke (1996), the inverse base-rate effect of Medin and Edelson (1988) is best understood as a case of highlighting in which the differential order of learning happens to be driven by differential base rates; that is, the role of differential base rates is to cause the more frequent cases to be learned before the less frequent cases.

One implication of the results is that theories of learning must be sensitive to order or learning. The highlighting effect occurs because later learned outcomes are learned in the mental context of previously learned outcomes. Theories of learning that are insensitive to learning order will necessarily fail to account for highlighting.

One family of theories that is insensitive to trial order comprises current versions of Bayesian learning models that treat all instances as equally representative data, irrespective of when the trials occurred. In principle, Bayesian learning models are able to incorporate temporal variables, but most current models do not, merely for simplicity. Future Bayesian approaches should include mechanisms that are sensitive to trial order, and, even better, also incorporate attentional learning.

Rather than incorporate explicit learned temporal dependencies into a Bayesian model, a different approach to modeling highlighting is to ‘‘break’’ the Bayesian model so it becomes non-Bayesian. Daw et al. (2008) showed that by applying various restrictions to the Kalman filter (a Bayesian model), which were motivated by different statistical approximation techniques, the approximately Bayesian model could qualitatively reproduce the basic torsion of the highlighting effect. Some approximations yielded the basic torsion, while others did not. It remains to be seen whether any particular approximation to a Kalman filter can exhibit the full set of preferences reported in Table 2 and results from previous studies summarized in the first half of this chapter.

A theorist might be motivated to model highlighting with an approximately Bayesian model if highlighting is appraised as a mere breakdown in an otherwise smoothly operating Bayesian mind. But highlighting is not a mere anomaly, and highlighting is not dependent on straining the limits of human information processing. As was suggested in the first half of this chapter, highlighting is a robust phenomenon that occurs across a variety of situations and with only moderate demands on mental processing. Highlighting is a sign of learning well, not badly.

Instead of asserting that the mind is a poor approximation to a Bayesian model, the Bayesian theorist can maintain that the mind is Bayesian, but at different levels of analysis. Rather than insist that the entire mind is globally Bayesian, it may be that sub-processes or components of mind are locally Bayesian. Clearly there may be learning occurring at many levels: neurons, anatomical partitions of brain, functional partitions of mind, individual persons, committees of people, corporations, and entire societies. Locally Bayesian learning of cue-to-attention mappings and of attended-cue-to-outcome mappings is one candidate for this sort of approach (Kruschke, 2006c).

Another implication of the results presented here is that explanations of the inverse base-rate effect should start with an explanation of (canonical) highlighting, and then include additional considerations for dealing with response biases in the presence of differential base rates. In particular, because the preference for the later learned category, in response to probe PE.PL, persists even when base rates of the response categories oppose that preference, there is likely to be an underweighting of base rates (Johansen et al., 2007; Kruschke, 1996). Indeed, Goodie and Fantino (1999) argued that base rates are rationally underweighted because base rates change more often than cue-outcome contingencies (see Dunwoody, Goodie, & Mahan, 2005, for empirical evidence).

One further implication of the results is that highlighting is not caused by eliminative inference (Juslin et al., 2001). The theory of eliminative inference assumes that response L is given to cues PE.PL because the L category has not been well learned: The L response is given because the well-learned E response is eliminated, and category L is inferred because it is all that remains from the response options. In the canonical design, however, all the categories are very well learned; there are no rare categories that are only weakly learned (cf. Experiment 2 of Kruschke, 2001b).

It is hoped that the canonical design, data, and Bayesian posterior can be profitably applied by other researchers. The canonical design is adaptable to various stimulus types and subject populations. In particular, if the phases are trained to criterion, rather than for a fixed number of trials, genuine learning of the early cases is assured. The computer program that was used for the experiment is available from the author's Web site. Because the canonical, equal base-rate design is particularly challenging to theories of learning, the data can be used as a test bed for models of learning. The complete raw data set is available from the author's Web site. Analysis of future data sets, when using the hierarchical Bayesian model of Figure 5, might also profitably use the posterior beliefs from the present analysis to inform priors for subsequent analyses. The posterior distribution is also available from the author's Web site.

ACKNOWLEDGMENTS

For helpful comments on drafts of this chapter, I thank Rima Hanania and Brian Ross. For help administering the experiment, thanks go to Kaitlyn Smith, Kari Vann, and Phaedra Willson. Stephen Denton provided greatly appreciated assistance with maintaining the computers on which the experiment was run. For providing pilot data, I thank Young Ahn, Stephen Denton, Rima Hanania, and Richard Hullinger.

Appendix: Two general reasons that null hypothesis significance testing has less than ambient pressure

A traditional chi-square analysis might be applied to the data in Table 2, but there are compelling reasons to avoid null hypothesis significance testing (NHST) in favor of Bayesian analysis, in general. One reason to avoid NHST is that it relies on the covert intention of the experimenter to define what it means to replicate the experiment and thereby derive a sampling distribution. The sampling distribution for replicated experiments is the crucial foundation for NHST, because the sampling distribution determines the p value. In traditional NHST, a replication usually assumes that the intention was to fix the sample size N, whereby a replication of the experiment means a random sample of size N from the null hypothesis population. In the present experiment, N was not fixed in advance. Instead, available session times were posted for many hours during a week. Many volunteers signed up. If the supply of volunteers happened to be at a slow rate, the experiment would have been run for another week or two. Only a subset of those who signed up actually showed up for the experiment. On very rare occasions, a computer may inexplicably freeze during an experiment, or a subject might decide to discontinue the experiment. After the data are collected, the learning criterion excludes some unforeseen number of subjects from further analysis. It is absurd, therefore, to consider a sampling distribution in which N is fixed. But all the p values computed by statistical packages, and critical values tabulated in the appendices of textbooks, assume fixed N. More fundamentally, the experiment was designed to insulate the data from the experimenter's intentions, so the experimenter's intention to run $N = 20$ or $N = 200$ should have no influence on the interpretation of the data. The fundamental logic of NHST assumes that the experimenter's intentions *should* determine the interpretation of data, which runs counter to the even more fundamental effort to insure that the experimenter's intentions do *not* influence the data.

There is another reason to avoid NHST: It does not tell us what we want to know. Consider, for example, responses in the test phase to cue I. The data suggest that $p(E|I) > p(L|I)$ in the underlying population, but we would like to know how much we can believe that there is a difference. More generally, we would like to know how much we can believe in any particular difference $p(E|I) - p(L|I)$ in the underlying population. Suppose we conduct a chi-square goodness-of-fit test for a null hypothesis of equal response probabilities across the four response options. The resulting p value tells us the probability of getting a chi-square value as or more extreme than the one we found in our data, were we to repeat the experiment with the same N from the null hypothesis. The NHST p values tells us about the probability of data we might have gotten but did not observe, if we replicated according to covert intentions of fixed N. In principle, we could consider alternative hypotheses and conduct tests of rejection on those alternatives, to construct a range of underlying response probabilities that significance testing would not reject, but this is not done by standard statistical packages and would still rely, nevertheless, on the strange notion of a fixed-N sampling distribution. And, most importantly, it would not tell us how much we should believe in each unrejected set of response probabilities. Bayesian analysis, on the other hand, relies only on the observed data, not on the experimenter's intentions during data collection. And Bayesian analysis tells us what we want to know, namely, the believabilities of underlying response probabilities and their differences.

REFERENCES

Akgün, A. E., Byrne, J. C., Lynn, G. S., & Keskin, H. (2007). Organizational unlearning as changes in beliefs and routines in organizations. *Journal of Organizational Change Management*, *20*(6), 794–812.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Bohil, C. J., Markman, A. B., & Maddox, W. T. (2005). A feature-salience analogue of the inverse base-rate effect. *Korean Journal Of Thinking & Problem Solving*, *15*(1), 17–28.

Colunga, E., & Smith, L. B. (2008). Knowledge embedded in process: The self-organization of skilled noun learning. *Developmental Science*, *11*(2), 195–203.

Courville, A. C., Daw, N. D., Gordon, G. J., & Touretzky, D. S. (2004). Model uncertainty in classical conditioning. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems Vol. 16*, (pp. 977–984). Cambridge, MA: MIT Press.

Cramer, R. E., Weiss, R. F., William, R., Reid, S., Nieri, L., & Manning-Ryan, B. (2002). Human agency and associative learning: Pavlovian principles govern social process in causal relationship detection. *The Quarterly Journal of Experimental Psychology Section B*, *55*(3), 241–266.

Cunha, M. Jr., Janiszewski, C., & Laran, J. (2008). Protection of prior learning in complex consumer learning environments. *Journal of Consumer Research*, *34*(6), 850–864.

Cunha, M. Jr., & Laran, J. (2009). Asymmetries in the sequential learning of brand associations: Implications for the early entrant advantage. *Journal of Consumer Research*, *35*, 788–799.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 431–452). Oxford, UK: Oxford University Press.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

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

Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology*, *50*, 131–138.

Dunwoody, P. T., Goodie, A. S., & Mahan, R. P. (2005). The use of base rate information as a function of experienced consistency. *Theory and Decision*, *59*(4), 307–344.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194.

Fagot, J., Kruschke, J. K., Dépy, D., & Vauclair, J. (1998). Associative learning in baboons (papio papio) and humans (homo sapiens): Species differences in learned attention to visual features. *Animal Cognition*, *1*, 123–133.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2005). The role of prediction in construction-learning. *Journal of Child Language*, *32*, 407–426.

Goodie, A. S., & Fantino, E. (1999). What does and does not alleviate base-rate neglect under direct experience. *Journal of Behavioral Decision Making*, *12*(4), 307–335.

Hayes, B. K., Foster, K., & Gadd, N. (2003). Prior knowledge and subtyping effects in children's category learning. *Cognition*, *88*, 171–199.

Johansen, M. K., Fouquet, N., & Shanks, D. R. (2007). Paradoxical effects of base rates and representation in category learning. *Memory & Cognition*, *35*(6), 1365–1379.

Juslin, P., Wennerholm, P., & Winman, A. (2001). High level reasoning and base-rate use: Do we need cue competition to explain the inverse base-rate effect? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 849–871.

Kalish, M. L. (2001). An inverse base rate effect with continuously valued stimuli. *Memory & Cognition*, *29*(4), 587–597.

Kalish, M. L., & Kruschke, J. K. (2000). The role of attention shifts in the categorization of continuous dimensioned stimuli. *Psychological Research*, *64*, 105–116.

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. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 3–26.

Kruschke, J. K. (2001a). Cue competition in function learning: Blocking and highlighting. Talk presented at the third international conference on memory, July 2001, Valencia, Spain, (Available from the author's Web site).

Kruschke, J. K. (2001b). The inverse base rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 1385–1400.

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

Kruschke, J. K. (2003a). Attentional theory is a viable explanation of the inverse base rate effect: A reply to Winman, Wennerholm, and Juslin (2003). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *29*, 1396–1400.

Kruschke, J. K. (2003b). Attention in learning. *Current Directions in Psychological Science*, *12*, 171–175.

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

Kruschke, J. K. (2006a). Learned attention. Presentation at the fifth international conference on development and learning. Indiana University, Bloomington, (Available at the author's Web site).

Kruschke, J. K. (2006b). Locally Bayesian learning. In R. Sun (Ed.), Proceedings of the 28th annual conference of the cognitive science society, (pp. 453–458). Mahwah, NJ: Erlbaum.

Kruschke, J. K. (2006c). 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 & Behavior*, *36*(3), 210–226.

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

Kruschke, J. K., & Bradley, A. L. (1995). *Extensions to the delta rule for associative learning.* (Indiana University Cognitive Science Research Report #141. Available at the author's Web site).

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.

Lamberts, K., & Kent, C. (2007). No evidence for rule-based processing in the inverse base-rate effect. *Memory & Cognition*, *35*(8), 2097–2105.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

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.

Medin, D. L., & Bettger, J. G. (1991). Sensitivity to changes in base-rate information. *American Journal of Psychology*, *104*, 311–332.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68–85.

Neal, R. M. (1996). *Bayesian learning for neural networks*. New York: Springer.

Nelson, J. B., & Callejas-Aguilera, J. E. (2007). The role of interference produced by conflicting associations in contextual control. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(3), 314–326.

Parish-Morris, J., Hennon, E. A., Hirsh-Pasek, K., Golinkoff, R. M., & Tager-Flusberg, H. (2007). Children with autism illuminate the role of social intention in word learning. *Child Development*, *78*(4), 1265–1287.

Pieters, R., Warlop, L., & Wedel, M. (2002). Breaking through the clutter: Benefits of advertisement originality and familiarity for brand attention and memory. *Management Science*, *48*(6), 765–781.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*, 819–865.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rosas, J. M., & Callejas-Aguilera, J. E. (2006). Context switch effects on acquisition and extinction in human predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 461–474.

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

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp. 1–34). Hillsdale, NJ: Erlbaum.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology: Comparative & Physiological Psychology*, *37B*, 1–21.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*, *4*, 3–18.

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E. J., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of Personality and Social Psychology*, *96*, 305–323.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

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

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

van Osselaer, S. M. J., & Janiszewski, C. (2001). Two ways of learning brand associations. *Journal of Consumer Research*, *28*(2), 202–223.

Wedell, D. H., & Kruschke, J. K. (2001). Consequences of competitive learning on social preference. Talk presented at the 42nd Annual Meeting of the Psychonomic Society, Orlando, November 15–18.

Winman, A., Wennerholm, P., & Juslin, P. (2003). Can attentional theory explain the inverse base rate effect? comment on Kruschke (2001). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1390–1395.

Yoshida, H., & Hanania, R. (2007). Attentional highlighting as a mechanism behind early word learning. In D. S. McNamara & J. G. Trafton (Eds.), Proceedings of the 29th annual meeting of the cognitive science society, (pp. 719–724). Austin, TX: Cognitive Science Society.