A JOURNAL OF THE ASSOCIATION FOR PSYCHOLOGICAL SCIENCE aps

CURRENT DIRECTIONS in Psychological Science

Prediction Errors and Attention in the Presence and Absence of Feedback

Andy J. Wills

University of Exeter

ABSTRACT—*Contemporary theories of learning typically assume that learning is driven by prediction errors—in other words, that we learn more when our predictions turn out to be incorrect than we do when our predictions are correct. Results from the recording of electrical brain activity suggest one mechanism by which this might happen; we seem to direct visual attention toward the likely causes of previous prediction errors. This can happen very rapidly—within less than 200 milliseconds of the error-causing object being presented. It is tempting to infer that if learning is driven by prediction errors, then little can be learned in the absence of feedback. Such a conclusion is unwarranted. In fact, the substantial learning that is sometimes the result of simple exposure to objects can also be explained by processes of directing attention toward the likely causes of previous prediction errors.*

KEYWORDS—*prediction error; attention; perceptual learning; categorization; contingency learning; blocking; exposure learning; EEG; ERP; eye tracking*

> Nothing fails like success because we don't learn from it.
>
> Attributed to Kenneth E. Boulding (1910–1993)

We appear to learn more about things for which we initially make incorrect predictions than we do about things for which our initial predictions are correct—the element of surprise seems conducive to learning. What are the mental processes that lead to this phenomenon, and what are the implications for situations in which there are no obvious external indicators of whether one has succeeded or failed?

Address correspondence to Andy Wills, School of Psychology, University of Exeter, Washington Singer Labs, Perry Road, Exeter, EX4 4QG, United Kingdom; e-mail: a.j.wills@ex.ac.uk.

PREDICTION AND LEARNING

The relationship between errors of prediction and learning is best illustrated by a short example. Imagine you are an allergist, trying to discover which of several foods (peas, carrots, chicken, ham) produce an allergic reaction in your patient. Figure 1a illustrates what you learn as a result of your investigation. Given the information in this figure, which do you think is more likely to cause an allergic reaction in the patient—chicken or ham?

The most common response to this question is "ham." This is intriguing because, in both cases, you have seen the patient eat the food in question and develop a rash. You have seen this happen an equal number of times for each of the two foods. So, what underlies the belief that ham is more likely to cause an allergic reaction? It cannot be attributed to differences in prior beliefs about chicken and ham, because the result is also found with entirely artificial stimuli (such as meaningless abstract shapes; e.g., Wills, Lavric, Croft, & Hodgson, 2007; see also Fig. 1b).

The phenomenon at work, known as *cue competition*, is widely observed in humans and other animals (Shanks, 1995) and is predicted by the hypothesis that we learn more from prediction errors than we do from prediction successes. A prediction error occurs when an event in the environment differs from our expectations. The concept of a prediction error roughly equates to the everyday concept of being surprised (as opposed to being wrong). We were not surprised that the patient developed a rash after eating peas and chicken, because we could already predict the presence of an allergic reaction on the grounds that he had eaten peas, which we already knew caused a rash. We therefore learned relatively little about the relationship between chicken and allergic reaction. In contrast, we were somewhat more surprised that the patient developed a rash after eating ham, and hence we learned more about the relationship between ham and allergic reaction.

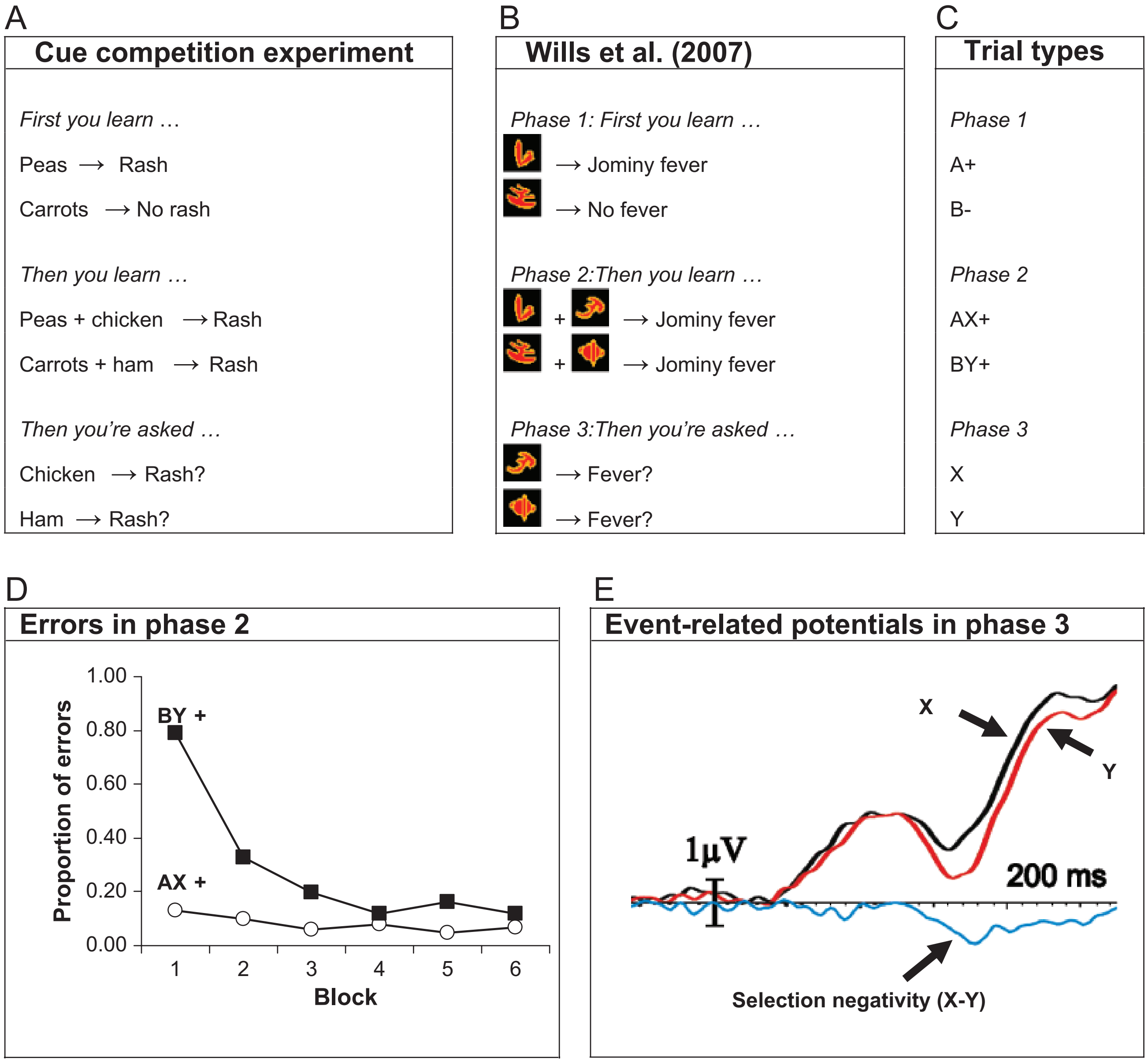

Fig 1. An illustrative cue-competition experiment (A) and a slightly simplified representation of the Wills, Lavric, Croft, & Hodgson (2007) cue-competition experiment (B, C) and its results (D, E). In the illustrative experiment (A), participants are told that a person eats certain foods and that consumption of these foods either does or does not cause them to develop a rash. Participants are then asked whether certain foods, not previously consumed in isolation (chicken, ham), will cause that person to develop a rash. In the Wills et al. (2007) experiment (Panel B), participants were asked to imagine they worked for a medical referral service and that their job was to predict a fictitious disease (''Jominy fever'') on the basis of ''cell bodies'' in patients' blood samples. These cell bodies were actually abstract shapes that were randomly allocated for each participant. On each trial, one or two cell bodies were presented, and participants made either a ''fever'' or a ''no fever'' response via key presses and received feedback on the accuracy of each response. Panel (C) shows this experiment expressed in the standard notation for learning experiments: Different letters (A, B, X, Y) indicate different stimuli, + indicates the presence of an outcome, – indicates the absence of an outcome. Proportion of prediction errors for AX+ and BY+ as phase 2 proceeds is shown in (D); a prediction error occurs when the participant incorrectly predicts the outcome of the trial. Panel (E) shows event-related potentials to the presentation of X alone and Y alone in phase 3. The third line shows the difference between the event-related potentials for X and Y—this difference is described as a selection negativity.

Associative theories propose that learning is the formation of associations between representations (for example, in the case of Pavlov's dogs, the formation of an association between a representation of a sound, and a representation of food). Classic associative-learning theory (e.g., Thorndike, 1898) is embarrassed by results such as cue competition, because the theory assumes that learning is simply driven by reinforcement. However, since the early 1970s, most associative theories of learning have incorporated the assumption that learning is driven by prediction error, largely on the basis of evidence such as cue competition and other related phenomena (see Pearce, 2008). These contemporary learning theories were primarily developed to explain animal learning, but from the early 1980s they became increasingly used to account for human learning as well (see Wills, 2005).

PREDICTION AND ATTENTION

One of the predictions of a number of contemporary associative theories is that differences in prediction error lead to differences in attention. For example, Pearce and Hall (1980) argue that learning is driven by prediction error because learners are limited in their ability to process the stimuli they encounter. To make best use of these limited resources, stimuli that have recently been followed by unpredicted events are prioritized by being given greater attention. This attentional differentiation leads to greater learning about stimuli that have recently been followed by unpredicted events.

In a recent study (Wills et al., 2007), we investigated the relationship between attention and prediction error in humans by measuring event-related potentials (ERPs) in a cue-competition experiment. ERPs are one way of estimating how an individual's electroencephalogram (EEG) changes in response to particular events. In our case, the event of interest was the presentation of an abstract shape that either had or had not previously been involved in substantial errors of prediction.

The design of our experiment is illustrated in Figure 1b. Participants were asked to imagine their job was to predict a fictitious disease on the basis of "cell bodies" in patients' blood samples. These cell bodies were actually abstract shapes. On each trial, one or two cell bodies were presented, and participants made either a "fever" or a "no fever" response via key presses and received feedback on the accuracy of each response. Figure 1c re-expresses the design of our experiment in the standard notation of learning experiments: Different letters (A, B, X, Y) indicate different stimuli, + indicates the presence of an outcome, and − indicates the absence of an outcome.

As Figure 1d shows, AX+ trials resulted in few prediction errors while BY+ trials initially resulted in many prediction errors. Behaviorally, our experiment showed the standard cue-competition result: When Y was subsequently presented alone, the probability of a participant responding "fever" was 0.72, whereas it was 0.45 for X presented alone. The stimulus previously involved in more prediction errors (Y) was more strongly associated with fever.

The phenomenon of cue competition is already well established behaviorally. The novel aspect of our study was to make use of the large body of knowledge about ERPs to identify "signatures" of attentional processing. In particular, given that our stimuli are distinguishable primarily by shape, one would expect to see a *selection negativity* if stimulus Y is attended to more than stimulus X (Hillyard & Anllo-Vento, 1998). In other words, one would expect the ERP for Y to be temporarily less positive than the ERP for X. This is what we observed, starting approximately 140 milliseconds after the onset of the stimulus (Fig. 1e).

There are two interesting implications of this result. First, it provides support for the idea that there is a correspondence between errors of prediction and the allocation of sensory processing. We may learn more from our prediction errors than from our prediction successes because the brain directs attention toward likely causes of previous errors. The second implication arises from the fact that the difference in attention is apparent less than 200 milliseconds after stimulus onset. It seems unlikely that much conscious deliberation is taking place during this very short amount of time.

PREDICTION IN THE ABSENCE OF FEEDBACK

Learning may be driven by errors of prediction, and one reason for this may be that objects involved in prediction errors attract attention. One might therefore be tempted to conclude that very little is learned in the absence of feedback. Such a conclusion would, however, be incorrect. Both humans and animals (see, e.g., Gibson & Walk, 1956) learn in the absence of feedback, as demonstrated by exposure tasks. Although people are not able to predict an outcome or category label during exposure tasks (because no such information is provided), they are able to make predictions about which aspects of the patterns tend to occur together. Errors in making these predictions can then drive learning about these co-occurrences.

One demonstration of exposure learning in humans can be found in our recent work (Wills, Suret, & McLaren, 2004). In the final part of our experiment, people had to learn to divide abstract patterns into two categories (see Fig. 2a). The task is not easy but can be done on the basis of overall similarity. This part of the experiment involved standard, feedback-informed learning: Participants were asked to guess which category each pattern belonged to and were told after each guess whether they were right or wrong. They performed poorly at first, but eventually they could categorize the patterns with a high degree of accuracy.

The key manipulation in this study involved comparing these participants to a different group that had been pre-exposed to the patterns before categorizing them. Pre-exposure involved asking this group of participants to look through a large number of patterns one at a time and say whether they had seen that particular pattern before (each pattern was presented exactly twice). They were not told whether their responses were right or wrong. These people then went on to learn to categorize the patterns in the presence of feedback. The people who had been pre-exposed to the patterns learned the categorization more quickly than did those who had not been pre-exposed. It is therefore clear that something about the stimuli was learned during pre-exposure. In other words, significant learning occurred in the absence of feedback.

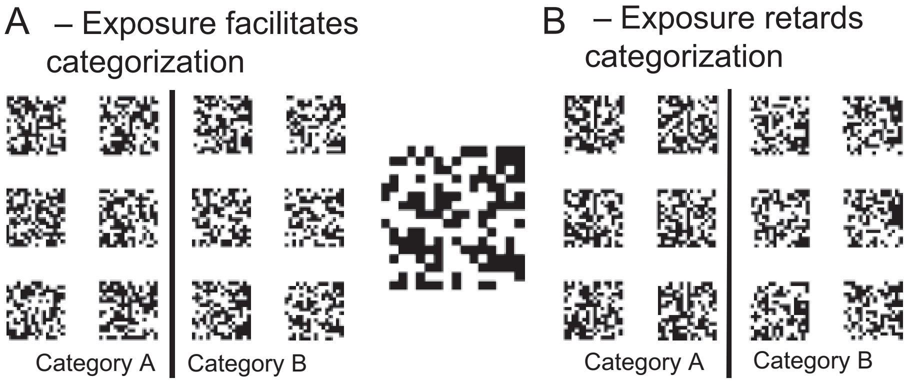
A – Exposure facilitates categorization
B – Exposure retards categorization
Category A
Category B
Category A
Category B

Fig. 2. An example of a categorization that is easier to learn if you look through a large number of the patterns first (panel A) and one that is made more difficult by pre-exposure (panel B). In our experiments (Wills et al., 2004), participants had to learn to divide the abstract patterns into two categories, as illustrated (although the number of patterns used in each case was much greater than 12). The opportunity to look through a large number of the patterns like those in panel (A) helped participants learn the categorization more quickly. In contrast, looking through the patterns in panel (B) resulted in the participants taking *more* time to learn the categorization. The critical, although nonobvious, difference between the two sets of patterns is that in A, aspects of the pattern that predict category membership are rarer than aspects of the pattern that do not. The pattern in the center of this figure is for illustration only—it is a magnification of one of the other patterns

Although this might seem to be a very different sort of learning than that discussed in the first part of this article, the existence of exposure learning can be predicted from the same central concepts; namely, (a) prediction error drives learning, and (b) stimuli previously involved in prediction error attract attention; prediction errors, in this case, mean failures to predict which aspects of the patterns tend to co-occur. Under these two central concepts, attention will be directed towards the parts of the pattern that are hardest to predict by co-occurrences (and hence that have high prediction error). This would, under some circumstances, make it easier to categorize the patterns later. This is because the aspects of objects that are hardest to predict by co-occurrence are often those that are relatively rare, and these rare aspects often tend to predict category membership better than common aspects.

It is, perhaps, not immediately obvious that rarity and diagnosticity (i.e., the ability to predict category membership) will often be correlated. The clearest, although somewhat extreme, example of rarity correlating with diagnosticity involves features that are present in all presented stimuli. Such features are clearly not rare (they always occur!), and, as a result of their ubiquity, they cannot provide a basis for dividing the stimuli into categories. Less common features can. Generally, to the extent that natural categories can be conceived as overlapping sets of features, there will be a correlation between rarity and diagnosticity. For example, the categories lion, tiger, and leopard all share a number of features (e.g., four legs). These features are therefore common across this set of three categories but are not diagnostic of lion versus tiger versus leopard. In contrast, stripes are relatively rare across this set of three categories, and they are highly diagnostic. However, in the lab it is possible to decouple rarity and diagnosticity. If one uses categories for which rarity and diagnosticity are not related, then it should be possible to abolish, or even reverse, the effect of pre-exposure. This is because in stimuli for which rarity and diagnosticity are unrelated, attention will not be selectively drawn to the diagnostic features of the stimuli (under the hypothesis, discussed above, that rarity and prediction error tend to be related).

As we predicted, in stimuli such as those shown in Figure 2b, for which rarity and diagnosticity are unrelated, pre-exposure leads to worse categorization performance than no exposure at all. The fact that two sets of patterns, which seem basically quite similar (Fig. 2a vs. Fig. 2b), can lead to opposite effects of pre-exposure seems quite counterintuitive. Nevertheless, it is predicted by an account that people make predictions about the co-occurrence of features and that those features that are hardest to predict by co-occurrence are the ones that become psychologically most salient.

CONCLUSIONS AND FUTURE DIRECTIONS

Effects such as cue competition suggest that we learn more when our predictions are incorrect than we do when our predictions are correct. Our recent study of ERPs supports the idea that one mechanism behind this phenomenon is the brain's tendency to direct attention toward those aspects of the environment that are most likely to have caused previous prediction errors. This direction of attention can be rapid (less than 200 milliseconds). The applicability of the idea that attention follows prediction error extends beyond situations in which feedback is obviously present—it can also be successfully applied to situations in which different parts of a presented object co-occur to some extent. Such situations permit learning to occur through simple exposure to objects. The concept of attention following prediction error can be used to predict when such exposure will help you to later categorize these objects and when it will hinder you.

In summary, the phenomena we observe can be explained by the two following principles: (a) Learning that is associative in nature is driven by prediction error, and (b) stimuli previously involved in prediction error attract attention. Such an account has a number of strengths. First, it is amenable to expression in formal mathematical terms (see, e.g., McLaren & Mackintosh, 2000), which allows clarity and specificity. Second, it has substantial generality across different learning tasks, including both those that include explicit feedback (Wills et al., 2007) and those that do not (Wills et al., 2004). Third, the account has generality across a variety of species (see Pearce, 2008, for a discussion of this account as it applies to nonhuman animals). Fourth, it makes predictions that are nonintuitive but nevertheless correct (for example, that exposure to the patterns in Fig. 2a helps you categorize them, but exposure to the patterns in Fig. 2b hinders categorization). More generally, studies such as those discussed here suggest that the brain seems to have a kind of "heads up" system that draws attention rapidly toward those aspects of our environment that are the likely causes of previous prediction errors and therefore may merit closer consideration. Although it is often said that we see what we expect to see, these studies suggest that we attend to that which has previously caused our expectations to be violated.

In this article, I have concentrated on prediction-error-driven associative learning. One might alternatively argue that humans learn equally about all of the relationships in the example in Figure 1a but that they then work through a series of reasoned inductive inferences. For example, when faced with the information *peas* + *chicken* → *rash*, participants might reason along the following lines: "Peas and chicken cause rash, but peas on their own cause rash; I therefore can't be sure that chicken caused the rash." When faced with *carrots* + *ham* → *rash*, however, they may reason as follows: "Carrots and ham cause rash, but carrots on their own do not cause rash; it therefore seems likely that the ham caused the rash."

One problem with such an account is that, although the attentional difference we observed (Fig. 1e) is sufficient to account for the cue-competition effect, it seems unlikely that this attentional difference results directly from conscious inferential reasoning, because it happens so quickly. One could also perhaps argue that the attentional difference is caused more indirectly by a conscious reasoning process modulating attention in a top-down manner. Measurements of gaze duration (Kruschke, Kappenman, & Hetrick, 2005; Wills et al., 2007) make such an account problematic: People spend more time looking at objects that have been involved in many prediction errors than at objects that have not, and this difference in gaze duration appears to be established before people become proficient at avoiding errors. If the attentional difference were the result of a top-down influence from a reasoning process, then one might expect it to follow, rather than precede, competent performance on the task.

Nevertheless, it is likely that learning sometimes does involve processes that cannot be captured by a simple associative system and that these processes may be more properly accounted for by a process of deliberative reasoning. There are phenomena that are problematic for associative accounts (e.g., De Houwer & Beckers, 2002) and for reasoning-based accounts (e.g., Le Pelley, Oakeshott, & McLaren, 2005), and a number of researchers have proposed multiprocess accounts of learning (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994). Examining the relative contributions of reasoning-based and associative processes, and the ways in which they interact, is an important direction for future research.

Recommended Reading

Ashby, F.G., & Maddox, W.T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 115–178. A clear summary of the idea that categorization is best understood as the result of multiple, competing brain systems.

Pearce, J.M. (2008). (See References). A well-written, very accessible introduction to animal cognition in general, including good coverage of associative theory.

Wills, A.J. (2005). (See References). An edited collection of relatively accessible articles on human associative learning.

Acknowledgments—Some of the research reported in this article was supported by a BBSRC grant 9/S17109. The writing of the article was supported by an EC Framework 6 project grant 516542 (NEST).

REFERENCES

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

De Houwer, J., & Beckers, T. (2002). Higher-order retrospective revaluation in human causal learning. *Quarterly Journal of Experimental Psychology*, *55B*, 137–151.

Erickson, M.A., & Kruschke, J.K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Gibson, E.J., & Walk, R.W. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology*, *49*, 239–242.

Hillyard, S.A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences, USA*, *95*, 781–787.

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

Le Pelley, M.E., Oakeshott, S.M., & McLaren, I.P.L. (2005). Blocking and unblocking in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*, 56–70.

McLaren, I.P.L., & Mackintosh, N.J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*, 211–246.

Nosofsky, R.M., Palmeri, T.J., & McKinley, S.C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

Pearce, J.M. (2008). *Animal Learning and Cognition*. Hove, UK: Psychology Press.

Pearce, J.M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552.

Shanks, D.R. (1995). *The psychology of associative learning*. Cambridge, UK: Cambridge University Press.

Thorndike, E.L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review Monograph Supplements*, *2* (Whole No. 8).

Wills, A.J. (2005). *New Directions in Human Associative Learning*. Mahwah, NJ: Erlbaum.

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

Wills, A.J., Suret, M.B., & McLaren, I.P.L. (2004). The role of category structure in determining the effects of stimulus preexposure on categorization accuracy. *Quarterly Journal of Experimental Psychology*, *57B*, 79–88.