Dissociable Learning Processes, Associative Theory, and Testimonial Reviews: A Comment on Smith and Church (2018)

Andy J. Wills · Charlotte E. R. Edmunds · Mike E. Le Pelley · Fraser Milton · Ben R. Newell · Dominic M. Dwyer · David R. Shanks

Abstract Smith and Church (2018) present a "testimonial" review of dissociable learning processes in comparative and cognitive psychology, by which we mean they include only the portion of the available evidence that is consistent with their conclusions. For example, they conclude that learning the information-integration category-learning task with immediate feedback is implicit, but do not consider the evidence that people readily report explicit strategies in this task, nor that this task can be accommodated by accounts that make no distinction between implicit and explicit processes. They also consider some of the neuroscience relating to information-integration category learning, but do not report those aspects that are more consistent with an explicit than an implicit account. They further conclude that delay conditioning in humans is implicit, but do not report evidence that delay conditioning requires awareness; nor do they present the evidence that conditioned taste aversion, which should be explicit under their account, can be implicit. We agree with Smith and Church that it is helpful to have a clear definition of associative theory, but suggest that their definition may be unnecessarily restrictive. We propose an alternative definition of associative theory and briefly describe an experimental procedure that we think may better distinguish between associative and non-associative processes.

Keywords categorization · classical conditioning · connectionist models · metacognition

A. J. Wills
School of Psychology, Plymouth University, U.K.
E-mail: andy@willslab.co.uk

C. E. R. Edmunds
Warwick Business School, Warwick University, U.K.

M. E . Le Pelley · B. R. Newell
School of Psychology, University of New South Wales, Australia.

F. Milton
Psychology, Exeter University, U.K.

D. M. Dwyer
School of Psychology, Cardiff University, Wales, U.K.

D. R. Shanks
Department of Experimental Psychology, University College London, U.K.

Summary of the target article

We begin with a brief summary of Smith and Church's (2018) article. The authors start out by noting that associative learning has long been a dominant framework in comparative psychology, but that cognitive psychologists often study situations that seemingly transcend associative learning, instead implicating explicit, conscious, cognitive processes. Their article is concerned with how to bring these ideas together: how best to characterize the relationship (and differences) between associative and cognitive processes.

Smith and Church first illustrate this idea via the phenomenon of uncertainty responses in non-human animals; that is, responses that allow an animal to terminate the current experimental trial in (for example) a perceptual discrimination task. These responses are referred to as "uncertainty" responses because they are more likely to occur on difficult trials than on easy trials, and hence are taken as evidence of a metacognitive state of uncertainty when faced with a difficult discrimination. Smith and Church argue that some types of uncertainty response are problematic for an associative learning account but naturally fit an explicit metacognitive account (p. 1568).

Smith and Church then turn their attention to a particular kind of category-learning task—the information integration (II) task—which they argue instantiates a "strong and sustainable associative-learning construct" (p. 1568). This argument is based, in part, on their view that humans lack conscious access to their strategy in the II task and, in part, on the involvement of the basal ganglia in II learning. Smith and Church go on to argue that inserting a brief delay between response and feedback in the II task leads to the "disabling of associative learning", with participants switching over to their explicit-declarative system (p. 1571). They subsequently expand on this idea, arguing that delaying feedback has similar effects in classical conditioning—that classical conditioning with a brief delay between cue and outcome (*trace conditioning*) is explicit, while classical conditioning without a delay is associative/implicit.

In the final part of their article, Smith and Church argue for the importance of a clear definition of associative learning. Their definition involves concrete stimulus inputs, simple behavioral outputs, immediate feedback, and the involvement of the basal ganglia. They argue that, under their definition, small changes in the sequencing and timing of reinforcement can qualitatively change the character of learning from associative to explicit. They argue that this effectively provides an "off-switch" for the primitive associaitve learning system, and that selectively disabling associative learning in this way may encourage non-human animals to "raise their cognitive game to the explicit level" (p. 1579). They further argue that the ability to switch off the associative system may provide a useful tool in the context of development and education; for example by allowing mathematics instructors to suppress children's rote, automatic habits and instead teach at a "a higher, explicit, conceptual level" (p. 1580).

Overview of our Comment

We agree with Smith and Church (2018) that it would be helpful to have a clear definition of associative learning. Our central critique is that progress on this clear definition is less likely to be made if one considers only the evidence consistent with one's preferred definition. Our Comment has the following structure. In the next section, we consider Smith and Church's review of work on the information-integration category structure, including the effects of introducing feedback delay. Then we consider their review of the effects of feedback delay in classical conditioning. Next, we present some definitions and clarifications that are used in subsequent sections, before turning to Smith and Church's review

of work on uncertainty responses in non-human animals. We go on to counsel against Smith and Church's recommendations for comparative and cognitive psychology, and make some alternative recommendations. We favor a definition of associative theory based on the concept of non-recurrent network architectures, and briefly describe the Shanks-Darby experimental procedure (Shanks & Darby, 1998), in which human behavior is non-associative by this definition, while rat and pigeon behavior is associative. Finally, we contrast *testimonial* reviews with *balanced* reviews, and suggest that scientific progress is more likely to be assisted by the latter.

The information-integration category-learning task

In this task, participants classify simple abstract stimuli (see Figure 1A) into categories, and receive feedback on their responses. An optimal description of the category structure in this task (see Figure 1B) would be something like "category A items are more dense than they are large". However, because the two dimensions use different, non-commensurable units, such a description is considered difficult or impossible to verbalize in practice—for the statement "more dense than large" to make sense, density and size must be converted to some common unit of measurement.

Smith and Church argue that "humans seem to lack conscious access to how they perform category tasks of this kind. They cannot declare the content of their category knowledge to others" (p. 1569). However, they present no evidence in support of this claim. When awareness is measured in this task, participants readily verbalize multi-dimensional strategies (Edmunds, Milton, & Wills, 2015; Edmunds, Wills, & Milton, 2019). The strongest tenable version of Smith and Church's argument is thus that the strategies participants verbalize are not the ones they use. One might see such a claim as supported by the fact that the strategies people verbalize in this task are poorly related to the strategies indicated by standard *decision-bound* computational modeling of their categorization responses (Edmunds et al., 2015, 2019). However, this decision–bound modeling is flawed (Edmunds, Milton, & Wills, 2018). The flaw is sufficiently fundamental that the results reported in Smith, Boomer, et al. (2014) – a key paper in Smith and Church's review – can be modeled under the assumption that all participants in the information-integration task use easily–verbalizable multidimensional rules (Edmunds et al., 2018), see also Donkin, Newell, Kalish, Dunn, and Nosofsky (2015) for a related conclusion.
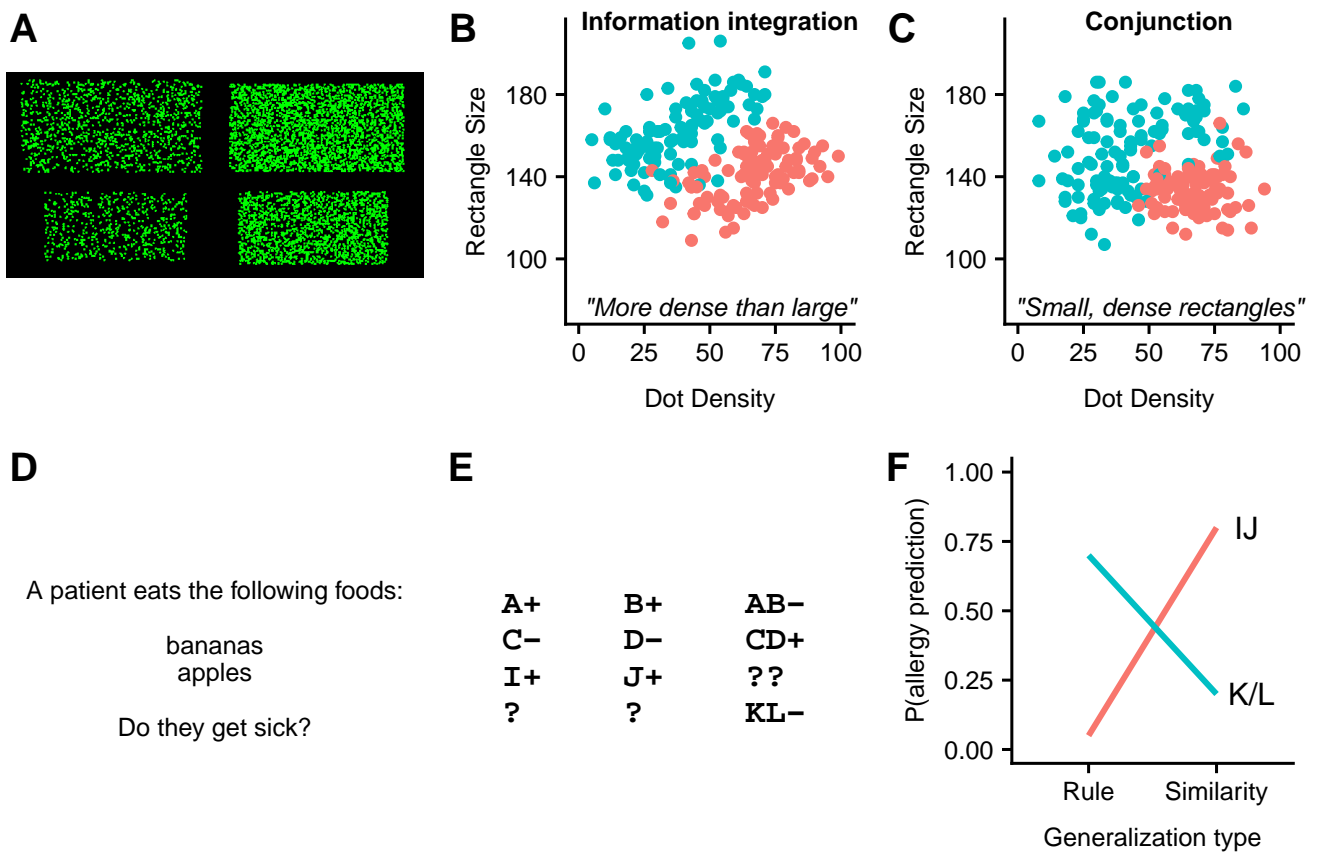
Fig. 1 A–C.: Category-learning task. **A.** Four example stimuli: rectangles of fixed aspect ratio, differing in size, and pixel density (see e.g. Smith et al., 2014). **B.** Information-integration category structure; each point represents one stimulus of a defined size and pixel density, color represents category membership. **C.** Conjunction category structure. **D–F.**: Shanks-Darby procedure. **D.** Example trial in a typical adult human procedure (procedure has also been used with different stimuli in rats and pigeons). **E.** Simplified design; letters represent foods, + = patient gets sick, - = patient feels fine, ? = untrained stimuli assessed at test. **F.** Rule- versus similarity-based generalization at test.

Smith and Church also discuss research concerning the insertion of a delay between response and feedback in the information– integration task (Maddox & Ing, 2005). We agree inserting such a delay can impair performance in category learning, but this effect is not restricted to information-integration category structures —similarly-sized effects can also be seen with category structures that are non–associative under Smith and Church's definition (Dunn, Newell, & Kalish, 2012). Thus, this evidence does not support the case that learning of infor-mation-integration category structures is qualitatively different from the learning of other category structures, or that immediacy of feedback critically determines whether learning is mediated by associative or explicit-declarative systems.

We also agree with much of Smith and Church's summary of the neural structures underlying information-integration category learning. However, similar neural structures are also involved in learning other category structures that, by Smith and Church's definition, are nonassociative. For example, a conjunction category structure (Figure 1C) can be optimally described by an easy-to-verbalize rule (e.g. "Category A items are small and dense"), and indeed this structure is considered to be a canonical example of a rule-based task by multiple-system theorists (e.g. Filoteo, Lauritzen, & Maddox, 2010). However, the brain regions involved in learning a (rule-based) conjunction category structure largely overlap with the regions involved in learning an infor-mation-integration structure (Carpenter, Wills, Bennattayallah, & Milton, 2016). Critically, this overlap includes the basal ganglia – a key component of Smith and Church's proposed associative learning system.

Hence, the relationship between the activation of particular brain regions and the involvement of specific mental processes is less clear than Smith and Church suggest. Further, to the extent there are differences in brain activation between conjunction and information-integration category learning, these differences seem con-

trary to Smith and Church's account. In particular, learning an information- integration structure leads to greater hippocampal activation than learning a conjunction structure (Carpenter et al., 2016). In contrast, Smith and Church argue that hippocampal involvement is characteristic of explicit, rule-based (non-assoc-iative) learning, and thus Carpenter et al.'s result seems inconsistent with Smith and Church's theory.

Classical conditioning

Aiming to expand their case about feedback delay beyond category learning, Smith and Church turn to evidence on classical conditioning in humans. They argue that classical conditioning with a brief delay between cue and outcome (*trace conditioning*) is explicit, while classical conditioning without a delay is implicit. They support this claim by citing work by Squire and colleagues (e.g. Clark & Squire, 1998), and use this work to reinforce their argument that brief feedback delays induce a switch from an implicit associative system to an explicit-dec-larative system.

Our first critique of this section is that it out of date—they cite 18 studies concerning the role of awareness in classical conditioning, but all are at least 17 years old. This is important because more recent research (e.g. Kattner, Ellermeier, & Tavakoli, 2012; Weidemann, Best, Lee, & Lovibond, 2013) includes failures to replicate early studies in this field, and demonstrates that in fact both types of conditioning require awareness (Lovibond, Liu, Weidemann, & Mitchell, 2011).

A second issue is that the available evidence on conditioned taste aversion seems incompatible with Smith and Church's case. In conditioned taste aversion, the delay between cue and outcome is typically long, sometimes several hours. Hence, conditioned taste aversion should be explicit by their definition—but this does not seem to be the case. For example, conditioned taste aversion in rats is still acquired when the illness-inducing outcome is delivered under general anaesthesia (Bermudez-Rattoni, Forthman, Sanchez, Perez, & Garcia, 1988; Samuel, Taub, Paz, & Raz, 2018). Also, in humans undergoing cancer treatment, conditioned taste aversions are formed despite the patients being aware that their illness/nausea is produced by the treatment and not the novel foods eaten (Bernstein & Webster, 1980).

What is associative learning?

In order to discuss Smith and Church's review of uncertainty responses clearly, it is important to first exam-ine their definition of associative learning more closely. For Smith and Church, *associative learning* is a theory of learning. They argue that the theory is a good one, which is productively incomplete. It is incomplete in the sense that it should not be expected to explain all learning, just a coherent subset thereof. Smith and Church seek to define that subset in terms of experimental procedure – there are specific procedural conditions under which an organism behaves in accordance with associative theory. They additionally seek to define the subset neurally – associative theory is applicable where certain brain structures, but not others, are involved. The textual analysis on which this summary is based is available in the Supplementary Materials.

We think there is a terminological issue with Smith and Church's definition; one that may lead to unintended confusion. Specifically, the term *associative learning* is also commonly used to mean a class of problem faced by an organism (see e.g. Mitchell, De Houwer, & Lovibond, 2009; Wills, 2005). Thus, associative learning is the class of phenomena in which an organism learns about a contingent relationship between events. A number of different theories seek to explain these phenomena; *associative learning*, in the sense Smith and Church use the term, is a term used to describe some of those theories. To reduce ambiguity, we will use the term *associative theory* in place of Smith and Church's term *associative learning* from here on.

More specifically, by *associative theory*, we mean a set of specific, formalized theories that together form a class. These theories are typically link-formation accounts, where experience of event pairings leads to strengthening or weakening of a connection between mental representations, in a way that can be captured in an algorithmic model. Canonical examples from the history of psychology include Hebbian learning (Hebb, 1949), the delta rule (Rumelhart, Hinton, & Williams, 1986), and several other associative accounts (e.g. Kruschke, 2001).

Membership of the class of associative theories is, like most categories, vague at the boundaries. However, we think most theorists would wish to exclude network models with recurrent connections from the associative theory class. To not do so would make it very difficult to reject the associative theory class, because some network models with recurrence are Universal Turing Machines (Siegelmann & Sontag, 1995), and hence can compute anything computable on a modern computer. Network models without recurrent connections are more limited in their computational power, and hence provide one meaningful boundary to the extent of the term associative theory. Thus, we define the

class of associative theories as those that can be instantiated in a non-recurrent network architecture.

Uncertainty responses in non-human animals

Smith and Church argue that while some un-certainty-response phenomena in non-human animals might be described associatively, to do so would be to stretch and strain the definition of associative theory unproductively. This conclusion, drawn previously by Smith, Couchman, and Beran (2014), has been previously challenged by multiple authors (see e.g. Basile & Hampton, 2014; Crystal, 2014; Le Pelley, 2014). Smith and Church do not consider these challenges in their review, so we reprise them briefly here.

Smith and Church's argument comes down to the idea that associative theory should restrict itself to physically present stimuli (a lever, a light, etc.), and hence learning involving (say) a memory (e.g. Hampton, 2001) cannot be associative. This definition of associative theory is at variance with decades of theoretical analysis (e.g. Mackintosh, 1983). More crucially, it excludes many standard conditioning phenomena, including cases studied by Pavlov, from the possibility of being accounted for by associative theory. For example, in a conditioned inhibition procedure (Pavlov, 1927), a tone is followed by food, but a compound of a tone and a light is not. In this situation, the animal seems to learn that the light prevents the delivery of food (as evidenced by, for example, the light's ability to reduce responding to another cue). On the standard associative account, conditioned inhibition occurs because presentation of the tone on tone-light trials evokes a memory of food, and this memory influences the learning that ensues. Simple associative theories that permit this kind of assumption provide a clear explanation of the findings of a wide range of metacognition studies, including the studies by Hampton (2001) and Shields, Smith, and Washburn (1997) that Smith and Church highlight. For an extended discussion of this point, the reader is invited to consult Le Pelley (2012).

Recommendations for comparative and cognitive psychology

In their final section, Smith and Church argue that small changes in the sequencing and timing of reinforcement qualitatively change the character of learning from associative to explicit. For the reasons we have outlined above, we are skeptical about this conclusion, and about the theoretical and applied potential of this technique.

Our conclusion echoes a number of previous debates in the literature. For example, the claims made by Smith and Church on the basis of feedback delay are similar in kind to claims made by others on the basis of concurrent load (Filoteo et al., 2010), observational training (Ashby, Maddox, & Bohil, 2002), and order effects (Spiering & Ashby, 2008). The first two of these phenomena have alternative explanations; see Newell, Moore, Wills, and Milton (2013) and Edmunds et al. (2015), respectively. The third doesn't seem to replicate (Edmunds et al., 2019). In all three cases, use of the information-integration category learning task was central to these experiments.

We suggest that it may be time to move away from the information-integration task to other, potentially more diagnostic, procedures. We agree with Smith and Church that one advantage of a clear definition of associative theory is that it facilitates our ability to study behavior beyond the scope of associative accounts. Given our earlier definition of associative theories as those based on non-recurrent network architectures, looking for behavior that requires recurrence to capture in a network model seems like a reasonable place to start. On that basis, rule-based generalization in the Shanks-Darby task (Shanks & Darby, 1998), see Figure 1D-F, is one promising option.[1] Rule-based generalization in this task is widely regarded as being beyond the scope of associative theory (see Wills, Graham, Koh, McLaren, & Rolland, 2011, for a discussion). In contrast, similarity-based generalization in this task can be captured by most associative theories. The majority of adult humans who learn the Shanks-Darby task to a high criterion show rule-based generalization (Shanks & Darby, 1998), while rats, pigeons, and adults under concurrent load, show similarity-based generalization (Maes et al., 2015; Wills et al., 2011). Rule-based generalization in this task leads to more frontal lobe activation than similarity-based generalization (Milton, Bealing, Carpenter, Bennattayallah, & Wills, 2017). We recommend the Shanks-Darby task as one promising procedure to investigate behavior beyond the scope of associative models.

Closing remarks

Smith and Church (2018) present what we describe as a *testimonial* review, by which we mean they present only the portion of the available evidence that is consistent with their theory, not reporting conflicting or contradictory evidence. An alternative approach is to present a *balanced* review, where both supportive and apparently contrary evidence are considered. A balanced review can still arrive at a clear conclusion, to the ex-

tent that the contrary evidence can be adequately addressed. We argue that the latter approach is more open, transparent and, crucially, more likely to assist scientific progress, than the former. By being clear about what is both supportive, and problematic, for a particular account, future research is more effectively directed to the questions that remain to be answered.

Note

1. Interested readers might also want to look at some other promising procedures discussed by McLaren et al. (2019).

Open practices statement

All the materials for this Comment are available at: https://osf.io/gvdxq/

References

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*, 666–677.

Basile, B. M., & Hampton, R. R. (2014). Metacognition as discrimination: Commentary on Smith et al. (2014). *Journal of Comparative Psychology*, *128*, 135–137.

Bermudez-Rattoni, F., Forthman, D. L., Sanchez, M. A., Perez, J. L., & Garcia, J. (1988). Odor and taste-aversions conditioned in anesthetized rats. *Behavioral Neuroscience*, *102*, 726–732.

Bernstein, I. L., & Webster, M. M. (1980). Learned taste-aversions in humans. *Physiology & Behavior*, *25*, 363–366.

Carpenter, K., Wills, A. J., Bennattayallah, A., & Milton, F. (2016). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*, 3557–3574.

Clark, R. E., & Squire, L. R. (1998). Classical conditioning and brain systems: the role of awareness. *Science*, *280*, 77–81.

Crystal, J. D. (2014). Where is the skepticism in animal metacognition? *Journal of Comparative Psychology*, *128*, 152–154.

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 933-948.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*, 840–859.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *Quarterly Journal of Experimental Psychology*, *68*, 1203–1222.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2018). Due process in dual process: Model-recovery simulations of decision-bound strategy analysis in category learning. *Cognitive Science*, *42*, 833–860.

Edmunds, C. E. R., Wills, A. J., & Milton, F. (2019). Initial training with difficult items does not facilitate category learning. *Quarterly Journal of Experimental Psychology*, *Published online*.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, *21*, 415–423.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 5359–5362.

Hebb, D. (1949). *The organization of behavior*. New York.

Kattner, F., Ellermeier, W., & Tavakoli, P. (2012). Both trace and delay conditioning of evaluative responses depend on contingency awareness. *Learning and Motivation*, *43*, 35–47.

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

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 686–708.

Le Pelley, M. E. (2014). Primate polemic: Commentary on Smith, Couchman, and Beran (2014). *Journal of Comparative Psychology*, *128*, 132–134.

Lovibond, P. F., Liu, J. C. J., Weidemann, G., & Mitchell, C. J. (2011). Awareness is necessary for differential trace and delay eyeblink conditioning in humans. *Biological Psychology*, *87*, 393–400.

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford, England: Clarendon Press.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 100–107.

Maes, E., De Filippo, G., Inkster, A., Lea, S. E. G., De Houwer, J., D'Hooge, R., ... Wills, A. J. (2015). Feature- versus rule-based generalization in rats, pigeons and humans. *Animal Cognition*, *18*, 1267–1284.

McLaren, I. P. L., McAndrew, A., Angerer, K., McLaren, R., Forrest, C., Bowditch, W., ... Verbruggen, F. (2019). Mackintosh lecture-: Association and cognition: Two processes, one system. *Quarterly Journal of Experimental Psychology*, *published online*.

Milton, F., Bealing, P., Carpenter, K. L., Bennattayallah, A., & Wills, A. J. (2017). The neural correlates of similarity- and rule-based generalization. *Journal of Cognitive Neuroscience*, *29*, 150–166.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*, 183-.

Newell, B. R., Moore, C. P., Wills, A. J., & Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves "implicit" perceptual categorization. A comment on Filoteo, Lauritzen and Maddox (2010). *Psychological Science*, *24*, 386–389.

Pavlov, I. P. (1927). *Conditioned reflexes*. London: Oxford University Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations*. Cambridge, MA: MIT Press.

Samuel, N., Taub, A. H., Paz, R., & Raz, A. (2018). Implicit aversive memory under anaesthesia in animal models: A narrative review. *British Journal of Anaesthesia*, *121*, 219–232.

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

Shields, W. E., Smith, J. D., & Washburn, D. A. (1997). Uncertain responses by humans and rhesus monkeys (Macaca mulatta) in a psychophysical same-different task. *Journal of Experimental Psychology: General*, *126*, 147–164.

Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, *50*, 132–150.

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*, 447-457.

Smith, J. D., & Church, B. A. (2018). Dissociable learning processes in comparative psychology. *Psychonomic Bulletin & Review*, *25*, 1565-1584.

Smith, J. D., Couchman, J. J., & Beran, M. J. (2014). Animal metacognition: A tale of two comparative psychologies. *Journal of Comparative Psychology*, *128*, 115–131.

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information-integration but not rule-based category learning. *Psychological Science*, *19*, 1169–1177.

Weidemann, G., Best, E., Lee, J. C., & Lovibond, P. F. (2013). The role of contingency awareness in single-cue human eyeblink conditioning. *Learning and Memory*, *20*, 363–366.

Wills, A. J. (2005). *New directions in human associative learning*. Hove: Lawrence Erlbaum Associates.

Wills, A. J., Graham, S., Koh, Z., McLaren, I. P. L., & Rolland, M. D. (2011). Effects of concurrent load on feature- and rule-based generalization in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*, 308-316.