COL·LECCIÓ
HOMENATGES 51

Associative Learning and Cognition

Homage to Professor N. J. Mackintosh. «In memoriam» (1935-2015)

Edited by
J. B. Trobalon, V. D. Chamizo

UNIVERSITAT DE BARCELONA

Associative Learning and Cognition

Associative Learning and Cognition

Homage to Professor N. J. Mackintosh. «In Memoriam» (1935-2015)

Edited by

J. B. Trobalon, V. D. Chamizo

UNIVERSITAT DE BARCELONA

Adolf Florensa, s/n
08028 Barcelona
Tel.: 934 035 430
Fax: 934 035 531
www.publicacions.ub.edu
comercial.edicions@ub.edu

ISBN 978-84-475-4056-3

Contents

Acknowledgments, by V. D. Chamizo, J. B. Trobalon 7
A few words, by Leonora Brosan Mackintosh 11
Preface. The Construct of Attention and Beyond: Homage to N. J. Mackintosh (1935-2015), by Ralph R. Miller 13

Geoffrey Hall, Mackintosh and Associationism 21
I. P. L. McLaren, K. Carpenter, C. Civile, R. McLaren, D. Zhao, Y. Ku, F. Milton, F. Verbruggen, Categorisation and Perceptual Learning: Why tDCS to Left DLPFC Enhances Generalisation..... 37
Dominic Michael Dwyer, Considering the Challenge of Mackintosh 2009: (Un)self-supervised Perceptual Learning? 69
G. M. Aisbitt, R. A. Murphy, An Application of a Theory of Attention (Mackintosh, 1975) to Psychopathy: Variability in the Associability of Stimuli..... 89
Janie Lober, Irina Baetu, A. G. Baker, Bottom-up Associative Mechanisms and Generalization Can Account for Apparent Contrast Effects Between Causes of Different Strengths 109
Paula J. Durlach, Alleviation of Acute Caffeine Withdrawal Reinforces Flavor Liking 141
Apolonia Manchón, Marta N. Torres, Teresa Rodrigo, V. D. Chamizo, Successive Contrast Effects in a Navigation Task with Rats 157
Anthony Dickinson, Instrumental Conditioning Revisited: Updating Dual-Process Theory..... 177
Richard A. Inman, Robert C. Honey, John M. Pearce, Asymmetry in the Discrimination of Auditory Intensity: Implications for Theories of Stimulus Generalisation 197
Gabriel Ruiz, Nicholas J. Mackintosh and the Renaissance of Animal Psychology in Spain: A Collaborative Enterprise 223

Appendix. Publications by Professor N. J. Mackintosh in Collaboration with UB members..... 253

Acknowledgments

The chapters published in this volume are a homage to Professor N. J. Mackintosh (1935-2015), an outstanding academic and a dear friend and colleague to all of the participants. The topics have been freely chosen by the authors. The fact that this book appears in a specific collection ("Homages") of the publishing section of the University of Barcelona deserves some explanation. Professor Mackintosh collaborated with different members of the Departament de Psicologia Bàsica (at present, Departament de Cognició, Desenvolupament i Psicologia de l'Educació), Universitat de Barcelona (UB), from the beginning of the 1980s until he passed away (on 8th February, 2015), after a brief illness. We were all devastated by the news. For many years our collaborative research aimed to see whether the spatial and the temporal domains share the same or similar conditions, basic effects, and mechanisms. Our results showed that many of the phenomena found in experiments on Pavlovian and instrumental conditioning and simple discrimination learning were also observed in our laboratory in experiments where rats were required to locate a goal by means of two or more distal landmarks. These phenomena included: blocking, overshadowing, latent inhibition, perceptual learning, and changes in attention to relevant and irrelevant cues (see the Appendix section of this volume for these references). Standard associative theories could explain all these phenomena (Rescorla & Wagner, 1972; Mackintosh, 1975; McLaren, Kaye, & Mackintosh, 1989; Pearce & Hall, 1980; Wagner, 1981). All the previous results are inconsistent with O'Keefe and Nadel's original proposal (1978) that locale learning (i.e., behaviour based on a representation of allocentric space, or cognitive map) occurs non-associatively, in an all-or-none manner, and that animals constantly update their cognitive map of their environment.

During the last years the main part of our collaborative research aimed to see whether male and female rats trained in a triangular shaped pool to find a hidden platform whose location was defined by two sources of information — one particular corner of the pool and a salient landmark positioned immediately above it — differed in their preferred mode of solution (geometry of

the pool cue vs. landmark cue) and also in the amount they learned about these two cues. Our results (see the Appendix section) are in agreement with previous findings showing that males and females do not always use the same strategies when solving spatial tasks (Williams, Barnett, & Meck 1990), age being a critical factor. In addition, males and females do not learn the same about the two sources of information. A biological origin?

Our collaboration with Professor Mackintosh began in 1982, on the occasion of a stay of V. D. C.* in the Department of Experimental Psychology (nowadays, Department of Psychology) as a postdoctoral student attending a course on "Animal Learning and Physiological Psychology", which was granted by the European Science Foundation (an ETP twinning grant). For many years Professor Mackintosh was a formal member of the UB research group *Learning and Cognition: A Comparative Approach* (www.gracec.info/). This collaboration produced a considerable number of contributions to meetings, publications, granted research projects, and research stays by different members of this group in the Department of Psychology at Cambridge University. During these years Professor Mackintosh was a frequent visitor to our university. It is worth mentioning his participation as an Invited Speaker in various courses organized and subsidized by different Catalan institutions (UB Department of Basic Psychology, UB Faculty of Psychology, financial help from the UB Chancellor, and UB Institute of Education Sciences — ICE abbreviation in Spanish), and in various PhD tribunals as an external examiner. However, his most important legacy to the UB was his outstanding contribution to the joint publications during those years that are included in the Annex section of this volume. (For some information about his distinguished career see Miller, 2016 and Hall, 2016 in this volume; visit also www.psychometrics.cam.ac.uk/about-us/directory/nick-mackintosh.)

As a sign of recognition and gratitude, Professor Mackintosh received posthumously (11 November, 2015) the highest honor, a Gold Medal, that the University of Barcelona can give to a person who is no longer with us. His widow Leonora Brosan Mackintosh, Lee, collected it in his name from the Chancellor of the University (Professor Dídac Ramírez), in a moving ceremony held in the Main Hall (the "Paraninfo") of the Historic Building (www.ub.edu/ubtv/video/acte-homenatge-prof-mackintosh). Several of his children (as well as other friends and colleagues) also attended the event, which was followed

* At the University of Barcelona since 1980.

by a formal lunch. At the ceremony, this book was announced to commemorate the first anniversary of the Gold Medal prize. In fact, when Professor Mackintosh died he was a candidate to be nominated Doctor Honoris Causa from the University of Barcelona. That process was automatically stopped because a requirement at our university is that the nominee personally attends a formal ceremony to collect the prize.

Thank you so much Nick!

The editors would like to express their deep gratitude to all the authors that have made this book possible. We should also like to express our gratitude to Lucy Mackintosh, for excellent language review in some chapters.

V. D. Chamizo, J. B. Trobalon
June 2016

References

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

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 102-130). Oxford: Oxford University Press.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Oxford University 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.

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

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5-47). Hillsdale: Erlbaum.

Williams, C. L., Barnett, A. M., & Meck, W. H. (1990). Organizational effects of early gonadal secretions on sexual differentiation in spatial memory. *Behavioral Neuroscience*, *104*, 84-97.

A few words

Before I met Nick I had never been to Spain. He found that hard to believe and took me with him as soon as he could. The first time that I went I was absolutely entranced, and could see just why he was so enraptured. For a man who loved fine weather, flowing wine, and flowing conversation, Spain was perfect. He loved the country and the culture, and was very positive about Spanish psychology. He was very pleased that people here are interested in the kind of work that he thought so important. For as long as I knew him he studied Spanish, and as some people will know, he always tried to introduce his talks in Spanish. It was very important to him as a way of showing his respect for his Spanish colleagues, but only they will know how well he succeeded!

Nick particularly loved Barcelona, although learning Catalan may have proved too much for him! He was very grateful for his continuing association with Professor Chamizo, and for her constant generosity to him, and I too am very grateful for her friendship and generosity.

I am sure that I speak for Nick when I say that he would give you his heartfelt wishes that Spanish psychology may continue to flourish.

Leonora Brosan Mackintosh
June 2016

Preface. The Construct of Attention and Beyond: Homage to N. J. Mackintosh (1935-2015)

Ralph R. Miller
Binghamton University, State University of New York, USA

This volume, edited by Professors Trobalon and Chamizo of the University of Barcelona, is dedicated to the memory of Professor N. J. Mackintosh (1935-2015) in honor of his contributions to our understanding of the basic principles of associative learning. Professor Mackintosh received his PhD in Experimental Psychology from Oxford University in 1963 under the supervision of Professor Stuart Sutherland. He went on to serve on the faculties of the University of Oxford, Dalhousie University, the University of Sussex, and finally the University of Cambridge where he served as head of department from 1981 to 2002. At various times he held visiting professorships at the University of Hawaii, Bryn Mawr College, the University of California at Berkeley, Yale University, the University of Pennsylvania, and the University of New South Wales.

Professor Mackintosh's scientific findings and insights were communicated through hundreds of insightful empirical and theoretical articles as well as a number of books, the four of greatest impact being:

1) Sutherland, N. S. and Mackintosh, N. J. (1971). *Mechanisms of Animal Discrimination Learning.*
2) Mackintosh, N. J. (1974). *The Psychology of Animal Learning.*
3) Mackintosh, N. J. (1983). *Conditioning and Associative Learning.*
4) Mackintosh, N. J. (1998; 2011). *IQ and Human Intelligence.*

The most widely cited of these volumes is *The Psychology of Animal Learning*, which is nearly encyclopedic and was reportedly written with minimal notes. Consistent with his monumental memory, in my conversations with

him he recalled aspects of some of my own experiments that I myself had forgotten decades after they had first appeared in print. Notably, the last of Professor Mackintosh's four books reflects a whole second research field in which he also distinguished himself and greatly influenced the assessment of intelligence in humans as well as practices in our schools. However, the present volume focuses on basic learning and cognition, as studied by Professor Mackintosh and the numerous researchers who were inspired by him and his work in this domain.

In addition to his own scientific contributions, an equally significant contribution of his remarkable academic career was his mentorship of innumerable students, postdoctoral fellows, collaborators, and the many visiting scholars who passed through his laboratories at Sussex and Cambridge. Why did so many researchers make the pilgrimage to Sussex and Cambridge to exchange views with Professor Mackintosh? The attraction was not only his depth of knowledge and broad interests, but his readiness to discuss the visitor's research, well-seasoned with his warm hospitality. Here I speak from first-hand experience, having been one of those pilgrims.

A further means through which Professor Mackintosh greatly influenced the study of basic learning processes was in his service as editor-in-chief of both *The Quarterly Journal of Experimental Psychology* and the *Journal of Experimental Psychology: Animal Learning and Behavior*. In his editorial capacity, he advised a whole generation of researchers with his famously constructive decision letters.

In his lifetime, Professor Mackintosh's scientific contributions were recognized through numerous awards including his being elected a Fellow of the Royal Society, and his receiving the Biological Medal and the President's Award from the British Psychological Society. The renown of his empirical and theoretical work gave him the visibility that allowed him to serve as the world's leading spokesman/advocate for the behavioral study of animal learning and cognition.

Professor Mackintosh's own research on learning was far-ranging. But he often returned to the construct of *attention*, which he typically narrowed to *associability* to avoid the excess baggage carried by the construct "attention". His focus on attention started with Sutherland and Mackintosh (1971) in which *selective attention* to attributes of a stimulus was directed by reinforcement and was a conserved quantity. Related to this theme, he published many empirical papers examining attention/associability as assessed in studies of

discrimination in animals. By 1975, his working model (Mackintosh, 1975) had evolved so that attention was no longer rigorously conserved, consistent with variation of arousal influencing the total amount of attention. But attention was still quasi conserved in that increases in attention to cues that best predicted impending biologically significant events were accompanied by decreases in attention to less valid cues that were present on the same training trial. It is worth noting that although Mackintosh invoked selective changes in attention to different cues (or attributes of cues), alternative accounts were proposed by others to address so-called attentional phenomena. In contrast to Mackintosh's mechanism of attention/associability changing as a function of experience, the formal construct of modifiable attention can be circumvented. This is most readily seen in the model of Rescorla and Wagner (1972), in which the associability of a cue is fixed and instead the subject's acquired behavior modulates subsequent learning about the cue. For example, in the visual modality subjects learn what to direct their gaze at, although this approach is more difficult (but not impossible) to apply in some other modalities, such as audition.

Professor Mackintosh's next steps forward in his continuing studies of attention are best seen in McLaren and Mackintosh (2000, 2002; also see McLaren, Kaye, & Mackintosh, 1989). Conventionally, perception and associative learning were viewed as independent sequential processes with perception preceding associative learning. However, the Gestalt psychologists viewed learning as a by-product of the laws of perception. In contrast, McLaren and Mackintosh viewed many perceptual phenomena as by-products of the laws of associative learning. The McLaren and Mackintosh model is a real-time, micro-elemental theory of learning aimed largely at uniting perception with traditional associative learning. The model emphasizes building percepts based on three processes: 1) excitatory within-compound associations between micro-elements that are presented together; 2) inhibitory within-compound associations between micro-elements not presented together but sharing companion micro-elements, and 3) decreased associability of non-reinforced micro-elements scaled to number of presentations along with increased associability of reinforced micro-elements scaled to number of presentations.

Professor Mackintosh's final theoretical statement concerning attention/associability is presented in Pearce and Mackintosh (2010). This model is a hybrid of his 1975 model, which attributed increased attention/associability

to predictive cues of high validity, and the Pearce and Hall (1980) model, which attributed increased associability to cues that are followed by surprising outcomes. As there are phenomena consistent with both mechanisms, Pearce and Mackintosh proposed ways in which these two processes could coexist.

Given the current emphasis in the literature on learning being driven by error reduction, it should be noted that Mackintosh (1975) used a local error reduction rule for learning similar to that of Bush and Mosteller (1951), in which the error that drives associative acquisition for any given cue is the difference between the outcome anticipated based on that cue and the outcome experienced. This contrasts with learning being driven by the total error reduction as assumed in the Rescorla and Wagner (1972) model, in which the error reduced in learning is the difference between the outcome anticipated based on all cues present on a given trial and the outcome experienced. In the Rescorla-Wagner model, total error reduction is the process responsible for cue competition, whereas in Mackintosh (1975) cue competition is produced by decreases in the associability of less valid predictors of the outcome. Later formulations such as Pearce and Mackintosh (2010) used a total error reduction mechanism. But Professor Mackintosh was not strongly committed to total as opposed to local error prediction. He retained the view that apparent total error reduction could be an artifact of changing attention and other processes (see Stout & Miller, 2007, for another account of cue competition that is not predicated on total error reduction). Professor Mackintosh was not theoretically rigid, in that he was willing to entertain a variety of somewhat contradictory models, recognizing that no contemporary model is comprehensive. Rather, he viewed models as heuristic devices to shape our thinking and direct us when designing experiments.

If there was one consistent feature of Professor Mackintosh's theorizing over the years, it was that he maintained a bottom-up orientation to his accounts of learning; that is, he consistently tried to explain so-called "reasoning" through simpler associative processes. More specifically, he viewed dyad associations as the foundation of many cognitive processes that others viewed as instances of top-down reasoning.

I close this preface by briefly describing five projects in my own laboratory that were significantly influenced by illuminating conversations I had with Professor Mackintosh over a fourty-year period. Any foolishness here is

of my own doing, but merit if any is in part a consequence of Professor Mackintosh's insights that greatly influenced my thinking and sometimes my experimental designs.

1) Implications of cue-to-consequence effects for stimulus associability: models that assign an associability to a given cue, whether it is variable as a function of validity or not, are unable to account for the now well documented superiority in learning some cue-outcome associations over other cue-outcome associations. Experiments have demonstrated that these cue-to-consequence effects cannot be explained simply by some cues and some outcomes having greater associability than other cues and outcomes (Garcia & Koelling, 1966; also see Foree & LoLordo, 1973, for an extension of this principle to responses). Seemingly, the only way to capture these effects in conventional models of learning is to posit an associability for the cue-outcome dyad (or cue-response outcome in the instrumental case studied by Foree and LoLordo). This problem challenges all contemporary models of learning (Miller & Grace, 2003).

2) Two quasi-independent types of memory interference: first, presentation of irrelevant stimuli near the time of training, or just prior or during a test, often disrupts acquired behavior, presumably because of competition for the limited capacity of working memory. The magnitude of these disruptions is directly related to the time between the disruptive irrelevant event and target training or testing, i.e., recency effects are observed here. Such interference can be couched in terms of the two memories, target and interfering, interacting because of their common time of activation. A second form of interference, often called associative interference, is often observed when a target association (Cue X-Outcome A) and a potentially interfering association have some but not all elements (e.g., Cue X-Outcome B or Cue Y-Outcome A; see Miller & Escobar, 2002). Our initial work contrasting these two types of interference (Miller, Greco, Marlin, & Balaz, 1985), one dependent on similarity in time and the other on similarity in non-temporal content, respectively, was greatly facilitated by conversations with Professor Mackintosh.

3) Over several decades, my collaborators and I developed a model of acquired behavior that emphasized information processing at the time of test (i.e., the comparator hypothesis, Miller & Matzel, 1988; Denniston, Savastano, & Miller, 2001; Stout & Miller, 2007), in contrast to most other models that focus on information processing during test (e.g., Mackintosh, 1975; Rescorla & Wagner, 1972). This orientation was diametrically opposed to that

taken by Professor Mackintosh in all of his published works, but he generously provided constructive advice to me over the several decades that we developed this model.

4) Professor Mackintosh and I shared reservations concerning whether accounts of learning based on reasoning and inference by the subject ever rose above simply relegating the question to a homunculus. One instance of learning in humans that is frequently explained in terms of inference is causal attribution. This prompted me to seek a demonstration of causal attribution in a nonhuman species, one presumably unlikely to employ complex inferences. Presumably, if rats exhibited behaviors analogous to behaviors that in humans are assumed to reflect causal learning, the same bottom-up associative processes that produced these behaviors in rats might also apply to causal learning in humans. Professor Mackintosh was highly supportive of this project, and in his role as the Devil's Advocate caused us to run additional experiments to assess alternative accounts of our rats' behavior that might have distinguished the behavior of our rats from that of humans exhibiting causal learning (Polack, McConnell, & Miller, 2013).

5) Perception learning: Professor Mackintosh devoted much of his efforts to examining the influence of learning on perception. The keystone to much of his later theorizing about this relationship was Espinet, Iraola, Bennet, and Mackintosh's (1995) demonstration of inhibitory perceptual learning. They found that many interspersed non-reinforced exposures to two compound stimuli, AX and BX, followed by pairings of B with an unconditioned stimulus (US) makes A inhibitory with respect to B or at least to the US. My colleagues and I developed a new technique to differentiate B being inhibitory with respect to A as opposed to the US, and found that indeed B was inhibitory with respect to A (also see Dwyer & Mackintosh, 2002). We also asked whether inhibitory perceptual learning obeys the same rules as conditioned inhibition. In conventional conditioning, few XB- trials followed by B-US trials produces excitatory sensory preconditioning of X; whereas many such trials makes X an conditioned inhibitor (Stout, Miller, & Escobar, 2004). Espinet et al. found that *many* AX/BX trials followed by B-US makes A (and X) inhibitory. We replicated this finding and additionally found that *few* AX/BX followed by B-US makes A (as well as X) excitatory. Hence, perceptual learning and conditioning obey parallel rules at least in this respect (manuscripts in preparation). In both of these projects, conversations with Professor Mackintosh greatly influenced our research.

The following chapters by a number of Professor Mackintosh's most prominent students, postdoctoral fellows and collaborators will give the reader a flavor of both the general focus of his primary interests as well as the diversity of specific questions that he and his collaborators have pursued. Professor Mackintosh's impact on the entire field of basic learning has been enormous.

Thank you, Nick. You will be missed.

References

Bush, R. R., & Mostellar, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*, 313-323.

Denniston, J. C., Savastano, H. I., & Miller, R. R. (2001). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 65-117). Hillsdale, NJ: Erlbaum.

Dwyer, D. M., & Mackintosh, N. J. (2002). Alternating exposure to two compound flavors creates inhibitory associations between their unique features. *Animal Learning & Behavior*, *30*, 201-207.

Espinet, A., Iraola, J. A., Bennett, C. H., & Mackintosh, N. J. (1995). Inhibitory associations between neutral stimuli in flavor-aversion conditioning. *Animal Learning & Behavior*, *23*, 361-368.

Foree, D. D., & LoLordo, V. M. (1973). Attention in the pigeon: The differential effects of food-getting vs. shock avoidance procedures. *Journal of Comparative and Physiological Psychology*, *85*, 551-558.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, *4*, 123-124.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London: Academic Press.

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

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

Mackintosh, N. J. (1998). *IQ and Human Intelligence*. Oxford: Oxford University Press.

Mackintosh, N. J. (2011). *IQ and Human Intelligence* (2nd ed.). Oxford: Oxford University Press.

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 102-120). Oxford: Clarendon Press / Oxford University Press.

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.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, *30*, 177-200.

Miller, R. R., & Escobar, M. (2002). Associative interference between cues and between outcomes presented together and presented apart: An integration. *Behavioural Processes*, *57*, 163-185.

Miller, R. R., & Grace, R. C. (2003). Conditioning and learning. In A. F. Healy & R. W. Proctor (Eds.), *Experimental psychology*, *Vol 4* (pp. 357-397), of *Handbook of Psychology*, I. B. Weiner (Ed.). New York: John Wiley & Sons.

Miller, R. R., Greco, C., Marlin, N. A., & Balaz, M. A. (1985). Retroactive interference in rats: Independent effects of time and similarity of the interfering event with respect to acquisition. *Quarterly Journal of Experimental Psychology*, *37B*, 81-100.

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

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: A review and a possible integration. In C. Mitchell, & M. Le Pelley (Eds.), *Attention and associative learning: From brain to behaviour* (pp. 11-39). Oxford: Oxford University Press.

Polack, C. W., McConnell, B. L., & Miller, R. R. (2013). Associative foundation of causal learning in rats. *Learning & Behavior*, *41*, 25-41.

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.

Stout, S. C., Escobar, M., & Miller, R. R. (2004). Trial number and temporal relationship as joint determinants of second-order conditioning and conditioned inhibition. *Learning & Behavior*, *32*, 230-239.

Stout, S. C., & Miller, R. R. (2007). Sometimes competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review*, *114*, 759-783.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Mackintosh and Associationism

Geoffrey Hall
University of York, UK, and University of New South Wales, Australia

Abstract. In his own estimation, N. J. Mackintosh's major contribution to psychology was to be found in his books. Foremost among these are the two (1974 and 1983) that dealt with animal learning and conditioning. The central theme of both books, implicit in the first and explicit in the second, was the way in which the phenomena of animal learning could be explained in terms of the notion of association formation (a review of the 1974 work referred to Mackintosh as "the compleat associationist"). In fact his associationism was not "compleat" — he is widely known for his emphasis on the role played by attentional processes in learning; and he was surprisingly modest in his assessment of the role of associative mechanisms in human learning. Nonetheless, an associative analysis was successfully applied not just to Pavlovian conditioning, but also to instrumental learning, avoidance, discrimination learning, spatial learning, and some aspects of perceptual learning. Although resisted by some, his persuasive writing established the associative account as the default position — that researchers today are still busy trying to prove him wrong is a tribute to the power and persistence of the ideas he developed.

Introduction

If I have given you delight
By aught that I have done,
Let me lie quiet in that night
Which shall be yours anon:

And for the little, little, span
The dead are born in mind,
Seek not to question other than
The books I leave behind.

R. Kipling, "The Appeal"

In an interview conducted shortly before his death, Nick Mackintosh was asked about which of his achievements (in psychology) he was most proud of.

TABLE 1. Mackintosh's authored books

Title	*Date*	*Citations*
Psychology of Animal Learning	1974	2338
Conditioning and Associative Learning	1983	1100
Mechanisms of Animal Discrimination Learning	1971	893
IQ and Human Intelligence	1998	625

Note: Citations are from Google Scholar, December 2015. In addition, Mackintosh edited three other books (discussed in the text), and a second edition of *IQ and Human Intelligence* was published in 2011. *Mechanisms of Animal Discrimination Learning* was co-authored with N. S. Sutherland.

He unhesitatingly identified his books; and although he did not appeal to us to restrict our questioning to these, he would no doubt have welcomed the fact that they continue to be worth our serious consideration (and sometimes, indeed, questioning).

Table 1 lists Mackintosh's authored books and also presents the number of citations of each, as recorded by Google Scholar at the end of 2015. His work on intelligence has been very influential, but, as he would himself allow, this was something of a sideline. His central contribution has been to the psychology of animal learning, and the citation counts for the first three books in the list attest to his importance in this field. In addition we should note that these three are book-ended by two edited works on the topic of learning. They were preceded by *Fundamental Issues in Associative Learning* (1969, edited jointly with W. K. Honig), and followed by *Animal Learning and Cognition* (1994), a volume constituting part of a *Handbook on Perception and Cognition*.[1] Both of these include lengthy essays by Mackintosh as editor. Taken together these works allow us to trace the development of Mackintosh's view of association formation as being a fundamental mechanism of cognition, and of his opinions concerning possible constraints on this analysis.

1. For completeness we should also mention his edited work (Mackintosh, 1995) on Cyril Burt.

Discrimination Learning and Attention

Although it appeared a couple of years after *Fundamental Issues*, the book written jointly with Stuart Sutherland, *Mechanisms of Animal Discrimination Learning*, deserves to be considered first. It had its origins in the work that Mackintosh did for his doctorate (Oxford, 1963). During the 1960s, attempts to establish a general theory of learning and behaviour were dominated by Hull's S-R (stimulus-response) reinforcement theory. The generation of psychologists trained at that time may have devoted their energies to finding fault with Hull, but they absorbed the general principle that association formation was a central explanatory concept. One influential figure at Oxford was J. A. Deutsch, who developed an account (e.g., Deutsch, 1964) that dealt with the phenomena considered by Hull, but which emphasized the formation of associations between the representations of stimuli. Another was Sutherland (Mackintosh's doctoral supervisor); he stayed with some form of S-R theory in his account of discrimination learning, but emphasized the role of perceptual or attentional processes in determining the nature of the S, and the role of learning in determining attention. Mackintosh's doctoral work was concerned with testing and developing this theory. It provided early instances of things that were to acquire substantial importance subsequently. At the empirical level there was an early demonstration of the phenomenon later known as *blocking* (Mackintosh, 1965); at the theoretical level it involved adoption of the notion that attention to a cue will be strengthened when the expectation of a particular outcome is confirmed, and will be weakened when it is disconfirmed — that learning depended on the ability of the cue to provide information.

When a full statement of the theory and of its supporting evidence appeared in the 1971 book, some aspects had already been overtaken by events. Mackintosh moved to Dalhousie University in 1967, and in 1968 helped organize the conference held there on which the *Fundamental Issues* volume was based. This conference included a report (by Kamin) of a version of blocking that was to become much more widely known than that of Mackintosh; and there were separate reports by Wagner (on the importance of cue validity) and from Rescorla (on correlational effects in conditioned inhibition) that supported the idea that learning about a cue depended on its informativeness or predictive power. In his final summary chapter Mackintosh sketched out how a theory, like that later proposed as the Rescorla-Wagner

(1972) model, could be developed, and might be able to deal with many of the phenomena of interest, without recourse to a concept of attention. But he also identified some phenomena that would be problematic for a theory of this sort. He discussed how an alternative theory, that gave a central role to changes in the properties of the stimulus (i.e., an attentional theory), might be developed, and outlined experiments that might be done to test it. Thus, even before the publication of the theory presented in *Mechanism of Animal Discrimination Learning*, Mackintosh was already anticipating the notions that would be expressed in his influential 1975 publication concerned with variation in stimulus associability (Mackintosh, 1975).

Publication of the 1975 theory was associated with a set of experimental reports testing its implications, and Mackintosh continued to study the way in which experience can modify stimulus processing, and hence discrimination, throughout his career. The outcome was a theory (McLaren, Kaye, & Mackintosh, 1989; see also McLaren & Mackintosh, 2000) that dealt with perceptual learning more generally, while retaining the essence of an associative account of conditioning. This last point is critical. Although we may accept that, in order to make the theory workable, it is necessary to specify the stimulus-processing mechanisms involved, it remains the case that the fundamental explanatory process is association formation — attentional and perceptual learning processes are subsidiary forms, adjuncts to the associative machine that some suppose to lie at the heart of cognition. Mackintosh's two most widely cited books were devoted to associative learning.

The Psychology of Animal Learning

For a previous generation, the "bible" on the topic of conditioning and learning had been the book of that title by Hilgard and Marquis (Kimble, 1961). The aim of the original work had been to set out the facts of conditioning so as to allow an assessment of the attempt to use conditioning principles to explain learning generally. It did not set out to summarise and compare rival theories; rather it laid out the facts in a systematic manner, noting their implications for theoretical positions as it went along. Mackintosh's *The Psychology of Animal Learning* took the same approach, and rapidly took the place of its predecessor. The sheer amount of experimental work that had been read, sorted, and digested, was impressive in itself. And the clarity of the presentation meant

that the outline of the whole of the wood could still be discerned even though individual trees were properly described. The student could rely on this new bible for instruction and guidance on the central issues and research on classical and instrumental conditioning, contrast, reinforcement, generalization, punishment — and so on. For many years this volume provided the starting point (if nothing more) for anyone wanting to learn about one of these topics (and it thus shaped the thinking of the generation brought up on it).

The book received an extensive review in the pages of the *Journal of the Experimental Analysis of Behavior* (Weisman, 1975), and the title given to this review clearly reveals what the book's central message was perceived to be: "The compleat associationist: A review of NJ Mackintosh's *The Psychology of Animal Learning*". Weisman's review acknowledged that there was a good deal in the book that was not advancing any particular theoretical line, but which consisted of empirical generalizations based on a thorough review of the relevant literature. Among the points picked out by Weisman for special mention are the following:

- That stimulus substitution theory still held its own as an account of the nature of the conditioned response.
- That classical CRs are not modified by their consequences.
- That instrumental training necessarily involves Pavlovian contingencies and these contribute to (in some cases, completely account for) the behaviour observed.
- That learning can occur about the relation between a response and its consequences, especially when performing the response generates salient feedback.
- That sensory preconditioning provides evidence that animals can associate motivationally neutral events.

But having noted such points Weisman went on to say that interwoven among this material was a "pervasive associationistic theory of learning". He expanded on this: "Mackintosh means that the neural correlates of events, stimuli, responses, and reinforcers are associated inside animals' heads...Animals learn what leads to what." (p. 386). Thus, he went on:

- In classical conditioning they associate stimuli with reinforcers.
- In instrumental learning they associate response with reinforcers.

- In punishment they learn to associate responses with aversive reinforcers.
- In avoidance they learn to associate responses with the omission of expected reinforcers.

In summary:

> Animals know about events after having learned correlations between them. And knowing is what cognitive psychology is all about. So the theory...is not just associationist but cognitive as well.[2]

Looking back at these remarks, from the perspective supplied by a distance of forty years, one is slightly puzzled that it was felt necessary to make them. It is a tribute to the power of Mackintosh's writing that the empirical points he made, and the theoretical line he took, are now so well established that we tend to take them for granted. That they initially provoked this sort of response from reviewers and other readers reminds us that, in his 1974 book, Mackintosh was still working at establishing what came to be the consensus. In fact he was somewhat ahead of his time, on two important issues.

First, the book outlines an account of the associative structures established by various conditioning procedures that is now widely accepted. This is a substantial achievement given the nature of the experimental evidence then available. We now have information from a range of clever experiments (conducted by, e.g., Holland, Rescorla, & Dickinson; for a review see Hall, 2002) that confirms the analysis offered by Mackintosh. But, for the most part this experimental work was done in the later 1970s (and later) and was not available to Mackintosh as he wrote.

Second is the emphasis given to the association as a central mechanism in explaining cognitive functioning. From one point of view there was nothing novel about this — indeed it is just what might be expected of a British psychologist brought up in the empiricist tradition of British philosophy. And given that the book is explicitly presented in the introductory chapter as an account of the development of the work pioneered by Pavlov and by Thorndike, their theoretical predispositions can be expected to show through (the latter, after all, was happy to describe himself as a "connectionist"). What is more novel is the almost total reliance on this explanatory mechanism. Reviewers

2. Weisman, 1975, p. 386.

upbraided Mackintosh for this. Weisman, in the review already discussed, complained that Mackintosh's approach led him to neglect areas of study, such as performance on various schedules of reinforcement, that are dear to the heart of those brought up in the atheoretical tradition of the experimental analysis of behaviour. (I suspect that Mackintosh's response would have been that there is no need to study such artificial contrivances — but that they would probably succumb to an associative analysis could one make the effort.) Another distinguished reviewer, J. A. Gray, objected to the failure of Mackintosh to endorse motivational explanations or to address motivational issues (Gray, 1975). Even in the case of avoidance learning (where, according to Gray, the case for a motivational explanation is strongest), Mackintosh insisted on an interpretation in terms not of conditioned motivational states, but of the development of expectations about the consequences of responding.

These are details; the important point about the analysis offered by Mackintosh is as follows. Conditioning studies are seen as a tool that can tell us about the association between particular event representations (sometimes called *nodes*); but the principles revealed by these studies will have relevance to the specification of a *conceptual nervous system* consisting of a huge array of such nodes corresponding to all perceivable stimuli (and possibly, all behavioural outputs). Psychological phenomena are assumed to be determined by the activation of these nodes, and behavioural adaptation by the formation of connections among them, and the propagation of activation around the network. These notions will now seem familiar, being those popularised rather later (e.g., by Rumelhart & McClelland, 1986) under the heading of *connectionism*; but they were anticipated by students of animal learning for whom Mackintosh spoke in *The Psychology of Animal Learning*.

Conditioning and Associative Learning

The theoretical approach that was implicit in *The Psychology of Animal Learning* was fully displayed nine years later with the publication of *Conditioning and Associative Learning*. The focus of the earlier book was its comprehensive review of the literature in various areas of learning; theoretical notions emerged as a consequence. The later book deals with much of the same empirical material, but now the focus is on explanation and mechanism, and the experimental findings discussed are just those that bear directly on theoreti-

cal issues. This book, like its predecessor, was the subject of a lengthy review in the *Journal of the Experimental Analysis of Behavior* (Williams, 1987). Much of Williams' review was concerned with a discussion of the distinction between Mackintosh's associative account and the Skinnerian approach likely to be favoured by most readers of that journal. But Williams also did us the service of summarising in ten fairly brief points, the major conclusions of Mackintosh's survey. I present a simplified synopsis of these points in Table 2. Again, as most of these points are so seemingly obvious and widely accepted, it is necessary to remind oneself that this was not so in 1983; that it is so now, is because of Mackintosh's work.

As Table 2 shows, the book was concerned almost exclusively with the analysis of classical and instrumental conditioning. The final chapter provided a foray into discrimination learning, but with the stated aim of showing (in the tradition of Spence, 1936) that this form of learning can be explained in terms of principles derived from simple conditioning. Other, possibly more complex, forms of learning were explicitly excluded from consideration (Mackintosh mentions, for example: problem-solving, imprinting, navigation, and performance on operant schedules of reinforcement), largely on the grounds that there was enough to say about conditioning. We can only speculate as to whether or not Mackintosh thought that these too could be explained in associative terms, given time and effort. A clue is provided in the very last paragraph of the book where he writes: "it should not be forgotten that animals are probably not just machines for associating events" (p. 277). Not much more is said; the only specific case cited is that of spatial learning, about which he writes:

> What does seem certain is that the perceptual processing and learning involved [in spatial learning] is somewhat more complex than anything involved in most studies of simple conditioning.[3]

Well if that is what he thought in 1983, I suspect that this was just because he had not yet got round to putting his mind to the subject. But he shortly did so; 1985 saw the publication of the first of a series of experimental studies (Diez-Chamizo, Sterio, & Mackintosh, 1985) on the topic of spatial learning, done in collaboration with Victoria Chamizo and colleagues at the Uni-

3. Mackintosh, 1983, p. 264.

Table 2. Williams' summary of *Conditioning and Associative Learning*

a)	Classical and operant conditioning are separate processes, involving different associative units; different rules of performance govern these conditioning processes.
b)	Stimulus substitution theory provides an adequate account of the nature of the conditioned response.
c)	The unit of learning in instrumental learning procedures is the response-reinforcer association; the function of the discriminative stimulus is to serve as a conditional cue informing the animal of the response-reinforcer relation.
d)	Punishment is the symmetric opposite of positive reinforcement – the animal learns the association between the response and the consequent aversive event.
e)	The effects of reinforcers involve two separate types of associations: between the response and the general hedonic effects of the consequent event, and between the response and the sensory properties of the particular reinforcer.
f)	Avoidance learning requires an analysis in terms of the events immediately consequent upon the avoidance response. Two-factor theory is unnecessary, because response-produced cues will become conditioned inhibitors with respect to aversive stimulation and thus assume positive value in their own right.
g)	Contingency effects can be derived from the more molecular principles that excitatory conditioning occurs whenever an "unpredicted" reinforcer occurs, and that inhibitory conditioning occurs whenever a "predicted" reinforcer fails to occur.
h)	The determinants of the strength of an association include not only the degree of temporal proximity between the elements of the association but other factors, including relative predictiveness, spatial contiguity, similarity, and "relevance".
i)	The degree to which a stimulus or response enters into an association depends upon past experience, as previous exposures of the stimulus or response in conditions in which nothing of consequence is predicted by those elements will cause them to lose "associability".
j)	Discrimination learning is best analyzed in terms of the more elementary processes of conditioning and extinction, as in the tradition of Spence.

Note: This is a simplified synopsis of the central points made by Williams (1987).

versity of Barcelona. The central issue was the extent to which spatial learning obeys the standard laws of associative learning, in contrast to the suggestion (as proposed by O'Keefe & Nadel, 1978) that it depends on the animal's ability to form some sort of spatial map or representation of its environment. This first paper manipulated intra-maze and extra-maze cues and demonstrated blocking and overshadowing effects like those seen in standard conditioning procedures. The authors concluded, modestly, that if spatial learning using extra-maze cues depends on the acquisition of some sort of map then the learning involved in this "interacts with other forms of learning in very much the same way as conditioning to a light interacts with conditioning to a buzzer" (p. 252). As the evidence began to build up over the course of the research programme that followed, modesty gave way to assurance. By 2002, Mackintosh was able to summarise his conclusions in a review paper with the unequivocal title: *Do not ask whether they have a cognitive map, but how they find their way about*. How they find their way about turns out to be interestingly complex — it certainly involves much more than a simple turn in response to a choicepoint stimulus, or the acquisition of approach strength by a cue located near the goal — but the mechanisms involved are based on, or at least consistent with, the principles of associative learning.

The successes of associative theory when skillfully applied in the field of spatial learning make one wonder about the other areas (mentioned above) that Mackintosh excluded from consideration. Could he have been equally successful in applying his general associative theory to these, if time (and inclination) had allowed? Sadly we will never know. We do know, however, of one area of psychology that he felt lay outside the scope of associative theory; I discuss this next.

Animal Learning and Human Cognition

The aim of animal learning theory is not to discover new facts about the behaviour of the laboratory rat; rather it is to devise an account of behaviour that has general relevance, applying, indeed, to our own species. This is what an earlier generation of psychologists (e.g., Hull, Skinner) claimed to have achieved. In his later writings, Mackintosh gives the impression of being distinctly embarrassed about such claims, and somewhat defensive about the scope of his own achievements. The introductory chapter of his 1983 book

refers to the claims of his predecessors as "extravagant" and acknowledges that, with the onset of a more "cognitive" psychology (a term to which we must return), people might be surprised that anyone should be continuing to study conditioning in the laboratory at all. He goes on to suggest, however, that such critics could be poorly informed — that modern learning theory is more complex and interesting, and has more to offer, than they know. This is certainly the tone of the introductory material for his final edited volume on the topic, *Animal Learning and Cognition* (1994), where he seems humbly grateful for the swing of the pendulum that has allowed a volume on animal learning to appear as part of a *Handbook of Perception and Cognition*.

The position is stated most clearly in a review article published a little later and entitled: *Has the wheel turned full circle? Fifty years of learning theory, 1946-1996* (Mackintosh, 1997). According to Mackintosh, at the start of this period learning theory occupied a central, even pre-eminent, position in psychology; but it was brought down low, as the wheel of history turned in the direction of cognition. This low point, about twenty-five years later, showed itself in two main ways. That many psychologists embraced the cognitive revolution had positive effects; there was a broadening of the range of topics studied — Mackintosh mentions categorization, concept learning, analogical reasoning, transitive inference, and several others. But there was a negative side: the risk that "in their haste to climb aboard a cognitive bandwagon, animal psychologists [might] abandon some hard-won achievements of classical learning theory" (p. 882). And a more serious problem was that psychologists generally, in their enthusiasm for cognitivism, were failing to recognize the real nature and scope of associative theory. Mackintosh's hope was that the wheel would turn again, if critics of learning theory could be more fully informed on these matters. It is worth quoting at some length his summary of his attempt to do this.

> Properly understood...associative learning theory is remarkably powerful. Of course, such a theory must acknowledge that the laws of association are much less simple than those of temporal contiguity between stimulus, response, and reinforcement. It must reject the restrictive assumption of S-R theory...and should assume that a representation of any event, be it an external stimulus or an action, can be associated with the representation of any other event, whether another external stimulus, a reinforcer, the affective reaction elicited by the reinforcer, or an animal's own actions. Equally important...it must allow that the representations

> of external events may be quite complex. They need not be confined to a faithful copy of an elementary association...they may be representations of combinations or configurations...once we have allowed associative learning theory these new assumptions we have a powerful account capable of explaining...behaviour that many have been happy to label cognitive and to attribute to processes assumed to lie beyond the scope of any theory of learning.[4]

The material that followed this statement, and put flesh on the bones of its claims, showed how the extended theory was capable of explaining performance on complex discrimination and categorization tasks, spatial navigation, and some instances of analogical reasoning. For these "appeal to more mysterious cognitive processes is often neither necessary nor helpful" (p. 890).

At this point the reader might be tending toward the conclusion that "cognition" is not to be regarded as a set of processes different from those assumed in associative learning theory, but rather, is simply a label for a set of phenomena that, it turns out, can be explained by means of the theory. But this is not Mackintosh's view. Having described the successes of associative theory he goes on immediately to say:

> Few psychologists, however, would deny the importance of a variety of cognitive processes when it comes to explaining our own behaviour...Associative analyses have an important role to play in any complete explanation of human behaviour [but] this is certainly not to deny the importance of numerous cognitive processes or operations that lie outside the scope of an elementary associative analysis. We do, for example, attempt to solve problems by inducing rules and testing our hypotheses. When we behave in this sort of rule-governed way, our behaviour is not amenable to a simple associative analysis.[5]

I can well imagine, in fact, that there will be many psychologists willing to dispute this intuition (see, for example, Skinner's, 1969, discussion of rule-governed and contingency-shaped behaviour in problem-solving). But the assertion relies on more than intuition; sceptics will need to deal with empirical bases for the claim, which includes observations like the following. When Mackintosh presented human subjects with discrimination tasks like

4. Mackintosh, 1997, pp. 883-884.
5. Mackintosh, 1997, p. 890.

those given to his pigeons, they normally showed rather different patterns of performance. They could be induced to behave like pigeons, however, in certain circumstances; for example, when the stimuli were too complex for a simple verbal description, and when a very rapid response was required. What could be more natural then, than to offer a dual-process account (see also McLaren et al., 2014) — a cognitive ruled-based system that operates under normal circumstances, and the associative system (held in common with other species) that comes into action when the other system is unable to function. Mackintosh went on to propose that the associative system was likely to be involved in a variety of implicit learning procedures; also that it provided a successful account of the performance shown when people are asked to judge contingencies between events (e.g., Dickinson & Burke, 1996). These are, no doubt, worthwhile achievements. But this reader, at least, is left with the feeling that this is something of a come-down for learning theory — to be assigned merely to a subordinate role, useful for dealing with events when you have no time to think.

Envoi

But we need not end on such a gloomy note. Human cognition may indeed involve processes other than those to be explained in terms of direct associations between the representations of events — but there is more to associationism than this. And it is all very well to describe these other processes as involving the use of rules or the manipulation of propositions, but these are indeed just descriptions rather than specifications of the mechanisms involved. The use of a rule or the construction of a hypothesis, are forms of behavioural adaption that are themselves in need of explanation. In one of his last contributions to the field, Mackintosh (in company with his collaborators; McLaren et al., 2014) made a start at dealing with this issue. I do not think he would have wanted to claim that the matter was settled, but we can record that the best attempt at explaining the nature of rule-based symbolic processing turned out to be in terms of a connectionist network; that is, in terms of a system using the associative principles that were central to his life's work in psychology.

References

Deutsch, J. A. (1964). *The structural basis of behavior*. Cambridge: Cambridge University Press.

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

Diez-Chamizo, V., Sterio, D., & Mackintosh, N. J. (1985). Blocking and overshadowing between intra-maze and extra-maze cues: A test of the independence of locale and guidance learning. *Quarterly Journal of Experimental Psychology, 37B*, 235-253.

Gray, J. A. (1975). Review of *The psychology of animal learning. Quarterly Journal of Experimental Psychology, 27*, 521-522.

Hall, G. (2002). Associative structures in Pavlovian and instrumental conditioning. In C. R. Gallistel (Ed.), *Stevens' handbook of experimental psychology* (3rd ed.) (Vol. 3, pp. 1-45). New York: John Wiley & Sons.

Kimble, G. A. (1961). *Hilgard and Marquis' conditioning and learning*. New York: Appleton-Century-Crofts.

Mackintosh, N. J. (1965). Incidental cue learning in rats. *Quarterly Journal of Experimental Psychology, 17*, 26-36.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London: Academic Press.

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

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

Mackintosh, N. J. (Ed.) (1994). *Animal learning and cognition*. San Diego, CA: Academic Press.

Mackintosh, N. J. (Ed.) (1995). *Cyril Burt: Fraud or framed?* Oxford: Oxford University Press.

Mackintosh, N. J. (1997). Has the wheel turned full circle? Fifty years of learning theory, 1946-1996. *Quarterly Journal of Experimental Psychology, 50A*, 879-898.

Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford: Oxford University Press.

Mackintosh, N. J. (2002). Do not ask whether they have a cognitive map, but how they find their way about. *Psicológica, 23*, 165-185.

Mackintosh, N. J., & Honig, W. K. (Eds.) (1969). *Fundamental issues in associative learning*. Halifax: Dalhousie University Press.

McLaren, I. P. L., Forrest, C. L. D., McLaren, R. P., Jones, F. W., Aitken, M. R. F., & Mackintosh, N. J. (2014). Associations and propositions: The case for a dual-process account of learning in humans. *Neurobiology of Learning and Memory, 108*, 185-195.

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibi-

tion. In R. G. M. Morris (Ed.) *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 102-130). Oxford: Clarendon Press.

McLaren, I. P. L., & Mackintosh, N. J. (2000). Associative learning and elemental representations. I: A theory and its application to latent inhibition and perceptual learning. *Animal Learning & Behavior*, *26*, 211-246.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

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

Rumelhart, D. E., & McClelland, J. L. (Eds.) (1986). *Parallel distributed processing, Vol. 1*. Cambridge, MA: MIT Press.

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. New York: Appleton-Century-Crofts.

Spence, K. W. (1936). The nature of discrimination learning in animals. *Psychological Review*, *43*, 427-449.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning*. New York and London: Academic Press.

Weisman, R. G. (1975). The compleat associationist: A review of N. J. Mackintosh's *The psychology of animal learning*. *Journal of the Experimental Analysis of Behavior*, *24*, 383-389.

Williams, B. A. (1987). The other psychology of animal learning: A review of Mackintosh's *Conditioning and associative learning*. *Journal of the Experimental Analysis of Behavior*, *48*, 175-186.

Categorisation and Perceptual Learning: Why tDCS to Left DLPFC Enhances Generalisation

I. P. L. McLaren, K. Carpenter, C. Civile, R. McLaren, F. Milton, F. Verbruggen

School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK

D. Zhao, Y. Ku

School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

Abstract. In the twenty-seven years that have passed since the McLaren, Kaye and Mackintosh (MKM) model of perceptual learning was first proposed, it has undergone considerable theoretical development and been subject to extensive empirical test. But we would argue that the basic principles of the theory remain as valid today as they were in 1989. One of these principles was that salience modulation of stimulus representations based on prediction error was a key component of latent inhibition and perceptual learning. It was this modification of what was otherwise a fairly basic adaptation of the model for categorisation proposed by McCleland and Rumelhart (M&R) that transformed a system that would exhibit enhanced generalisation between exemplars as category learning progressed, into one that would instead offer an improved capacity for discrimination between exemplars as a consequence of experience with the category. This modification has only been tested indirectly up until now, by looking at the predictions that flow from it and then comparing them to animal and human discrimination following stimulus pre-exposure. In this chapter we test this principle more directly, by using tDCS to disrupt the modulation of salience by prediction error, and show that when this is done, people exhibit the enhanced generalisation predicted by the standard M&R model. We conclude that our results provide further support for the MKM approach to stimulus representation.

Introduction

How we learn to distinguish between things is one of the basic questions for cognitive psychology. This paper focuses on two aspects of the mechanisms that allow us to do this. Categorisation in this paper refers to our ability to classify stimuli as members of one category or another as a result of trial and error training with members (exemplars) of the categories in question. Perceptual learning here refers to our enhanced ability to discriminate between certain stimuli as a consequence of experience with them or stimuli like them. Taken together, these two phenomena play a crucial role in learning to correctly identify stimuli as members of a particular class, and not confuse one stimulus with another similar one.

There are many theories and models of categorisation, and quite a few theories and models of perceptual learning. One of the few models that addresses both was originally proposed by McLaren, Kaye and Mackintosh (1989, henceforth MKM), in part as a response to and development of McClelland and Rumelhart's (1985, henceforth M&R) connectionist model of categorisation. It is this model that motivated the experiments discussed here, and, given the model-driven nature of our enquiry, we begin with a brief introduction to these models and the experimental paradigms we will use in this paper. We then go on to discuss how recent work using tDCS (trans-cranial Direct Current Stimulation) raises the possibility of influencing the error signal that drives learning and performance in the MKM model so as to change a participant's ability to distinguish between stimuli as a consequence of their experience with them. Our paper is an exploration of this possibility, and our results suggest both that perceptual learning and categorisation can be strongly influenced by anodal tDCS to frontal regions of the brain, and that a theory of perceptual learning and categorisation that relies on use of error-based modulation of the salience of the representations of stimulus input provides a good fit to the data we obtain using this preparation. We end by discussing the implications of these results for phenomena such as face processing.

Background

Two Models

McClelland and Rumelhart's seminal 1985 paper used the delta rule, an error correcting learning algorithm closely related to Rescorla-Wagner (Rescorla and Wagner, 1972), in a connectionist network employing distributed stimulus representations to model categorisation. We cannot do the model full justice here, but it was also noteworthy for its use of non-linear activation functions and a weight decay mechanism to help it produce both prototype and exemplar effects in what was effectively a single-layer (in that it has a single layer of modifiable weights) connectionist model. It did have one feature, however, that seemed to some of us problematic. This was that the learning algorithm coupled with the activation function inevitably led to units that were most frequently co-activated becoming more active as a consequence. This gave these units greater salience in later learning, and so it would be the units representing the more prototypical elements of a stimulus that would tend to form the strongest links to other units representing category membership.

This characteristic of the model may not be a problem for categorisation (though we will have more to say about this later) but it is certainly a problem for stimulus representation development as a consequence of experience with a category (i.e. for perceptual learning). McLaren, Leevers and Mackintosh (1994) were the first to show that humans trained to distinguish between two prototype-defined categories of stimuli (in this case chequerboards) were then actually better able to distinguish between two new exemplars drawn from one of these now familiar categories than between two exemplars taken from another entirely novel category that otherwise had a similar prototype-defined structure. The McClelland and Rumelhart model predicts the opposite result because, as illustrated in the lower half of figure 1, it will be the prototypical features contained within these new exemplars drawn from the familiar category that will be most salient. This will lead to the two exemplars being represented as more rather than less similar as a consequence, because these will tend to be the features shared by the two exemplars.

Our solution to this problem is shown in figure 1 (top half), which illustrates how the MKM theory predicts salience will change as a function of experience with exemplars of one category. The crucial difference between this

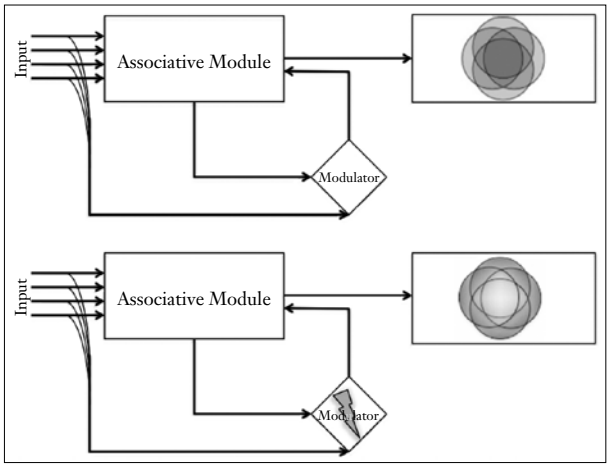

FIGURE 1. In both diagrams, each circle contains the set of elements representing one stimulus. Top half. This illustrates how modulation driven by prediction error (as in MKM) can be used to influence feature salience for a single category. The result, shown in the temperature diagram, is that stimulus features that are more predictable become less active (darker shading), leading to latent inhibition (slower learning as a consequence of pre-exposure). This improves discrimination between members of a prototype-defined category, as it relies upon the less predictable features unique to each stimulus. The bottom half of the figure shows how disrupting this modulatory input (e.g. using tDCS) reverses this effect (as in M&R), making the common, prototypical features of the stimuli the most salient (lighter shading).

model and that of McClelland and Rumelhart (1985) is that the activation of the units representing stimulus features (or elements, as we will often call them) is modulated by their error. Thus, if a unit is relatively unpredicted by other active units, but is externally activated because a feature corresponding to that unit has occurred and been perceived in the environment, then it's activation (i.e. salience for learning purposes) will be high because its error score will also be high. Conversely, if a unit receiving external input is well-predicted by other active units such that its error score is low, then its activation (and salience) will be low. This is exactly the opposite of the effect that occurs in the M&R model, and leads to new exemplars drawn from a famil-

iar, prototype-defined category being more easily discriminated, because the elements or features they share in common will be relatively low in salience, thus reducing stimulus similarity. The elements on which they differ (which will tend to be those elements that have changed from the prototype) will be relatively salient and this helps in learning to discriminate between them, as McLaren, Leevers and Mackintosh (1994) found.

The illustrations in figure 1 are similar to those of a figure in McLaren (1997) that is also used to explain how experience with exemplars drawn from a prototype-defined category will lead to better within-category discrimination. This 1997 paper, however, deals with one of the first reports of an analogue of the face inversion effect using artificial categories (again chequerboards) rather than faces. McLaren first trained participants to learn (by trial and error) to categorize chequerboard exemplars as belonging to one of two prototype-defined sets. The exemplars were made from the prototypes by randomly changing some of the black and white squares that made up the chequerboard that defined the category prototype, as in McLaren, Leevers and Mackintosh (1994). McLaren (1997) then demonstrated that an inversion effect could be obtained for new exemplars drawn from these now familiar categories, a result since replicated repeatedly by Civile, Zhao, Ku, Elchlepp, Lavric and McLaren (2014). The explanation for this result is that the exposure to the exemplars of the categories participants were trained on initially allows perceptual learning to take place as in McLaren et al. (1994), and this then improves discrimination and recognition performance to exemplars drawn from those categories that are in the usual upright orientation; but it does not help, and, as Civile et al. (2014) argue, actually hinders discrimination and recognition when these exemplars are inverted. This explanation depends on the MKM account of perceptual learning and categorisation, as the M&R model would once again predict the converse result.

There is thus some good evidence for the MKM modification of the M&R model of categorisation. We will use the categorisation followed by discrimination/recognition procedure just discussed later in this paper to test our hypotheses regarding the effects of frontal anodal tDCS stimulation on perceptual learning. But, before doing this, we first consider the prior issue of what tDCS might be able to offer in terms of influencing categorisation itself, and how tDCS might affect the type of error-based modulation that is the basis of the MKM model.

tDCS and Categorisation

Our first experiment investigates the effects of tDCS on a standard categorisation task that produces a prototype effect under normal circumstances (Posner and Keele, 1968). This work was inspired by the finding of Ambrus, Zimmer, Kincses, Harza, Kovacs, Paulus and Antal (2011), who provided evidence that tDCS could eliminate the prototype effect. There is other evidence that stimulation of PFC using tDCS can influence categorisation. Lupyan, Mirman, Hamilton and Thompson-Shill (2012) have produced some evidence that stimulation in frontal regions can enhance categorisation, and Kincses, Antal, Nitsche, Bártfai and Paulus (2003) have shown that when tDCS anodal stimulation was delivered over the left PFC (Fp3), probabilistic classification learning (PCL) was improved. Ambrus et al. (2011), however, found that anodal tDCS, applied to Fp3 during the training phase (and beginning 8 minutes before the training phase started) had a significant and quite different impact on categorisation performance in their version of the prototype distortion task. They obtained a significant *decrease* in performance accuracy in identifying prototype and low-distortion patterns as category members in the anodal group compared to the sham group. This is a striking aspect of their results as it is contrary to most studies that show increased performance when anodal tDCS is applied to task-relevant cortical areas during task execution (e.g., Fregni et al., 2005).

On close inspection, one possible interpretation of Ambrus et al.'s result is that anodal tDCS has reduced learning to the prototype, and increased generalisation to random patterns. This would have the effect of eliminating any prototype effect, and is exactly the type of pattern we would expect if the MKM model were to be transformed into the M&R version. Salience modulation enhances learning of novel stimuli, and so improves early acquisition of category discrimination, and it also reduces generalisation. Losing this type of modulation would lead to slower learning (at least initially) and greater generalisation. We speculated that anodal tDCS to Fp3 might have disrupted salience modulation by means of prediction error, leading to Ambrus et al.'s result.

If we now consider how tDCS might influence the brain's computation and use of prediction error, Reinhart and Woodman (2014) in a recent paper have shown that anodal tDCS over frontal regions can change prediction error. They used anodal stimulation at FCz and were able to show that this produced enhanced learning and selectively enhanced neural correlates of

prediction error. The most obvious conclusion to draw from this study is that 1.5 mA anodal stimulation applied with their electrode montage has the effect of amplifying prediction error, which will both speed learning and lead to the neural signature they found. This is not the effect we postulated in response to Ambrus et al.'s data, but it does suggest that prediction error can be influenced by anodal tDCS, and of course the locus of stimulation is rather different in Reinhart et al.'s work.

Our approach in the studies reported in this paper is to take something from the approaches of Ambrus et al.'s (2011) — because they were able to influence categorisation quite directly — whilst holding that of Reinhart et al. (2014) in mind — because they have good evidence for changing prediction error. Hence we employed a similar electrode montage to that used by Ambrus et al. (2011) stimulating Fp3, and increased the current from the 1 mA they used to 1.5 mA in the hope of maximising our chance of observing an effect on categorisation. If we were to observe such an effect, then we would consider the possibility that this effect would be due to our changing the contribution of prediction error in influencing learning and performance on the categorisation task. In this way we hoped to develop a procedure that would allow us to both influence categorisation and the perceptual learning that follows on from categorisation, which in turn would allow us to probe the mechanisms underlying both, using the MKM modification of M&R as our starting point for interpreting our results. Note that our procedure, which is akin to that used in earlier studies, employs electrodes (see later) that do not have a strongly focal effect, so that the stimulation we provide is perhaps best functionally described as left DLPFC rather than trying to claim any greater specificity.

Experiment 1

Experiment 1 is a conceptual replication of Ambrus et al. (2011), using a classic categorisation paradigm based on early work by Posner and Keele (1968) and Homa, Sterling and Treppel (1981) designed to reveal any prototype effect. We use three prototype-defined categories of chequerboards, with the exemplars in each category generated by adding noise (randomly changing a certain number of squares) of the prototype for that category. Participants are trained to classify exemplars into these three categories by trial and error, and then tested on exemplars and the prototypes (which are never shown in

training) to allow us to determine if an exemplar effect has occurred. Three types of stimulation, Anodal, Cathodal and Sham, are used, but all employ the same Fp3 electrode placement used by Ambrus et al.

Method

Participants

Fifty University of Exeter students (17 male) with a mean age of 21.5 years (sd 2.93) participated in the study. Two were excluded before analysis due to procedural complications, leaving 48.

Stimuli

These were 16 × 16 chequerboards containing approximately 50% black and 50% white squares. Four prototypes were created that were constrained to share 50% of their squares with one another, and also to consist of relatively clearly demarcated regions of black and white. This was achieved by making the colour of a given square depend on that of its near neighbours. Thus, if they were predominantly black then it was likely to be black, and vice-versa if the neighbours were predominantly white. Exemplars were generated by adding noise. A randomly chosen 96 squares would be set at random in a given prototype to generate an exemplar of that category, so that on average 48 squares are changed from the category prototype (see figure 2). In this way as many exemplars as were desired could be created. We used a total of 128 chequerboard exemplars from each of the four categories in these experiments, though not all of these stimuli would be used for a given participant. The stimuli used in the experimental phases (categorisation and test) were counterbalanced across subjects.

Participants were required to separate these chequerboard stimuli into three categories (A, B and C) during the training and test phases (see figure 2). In the training phase, 64 novel exemplars from each of the three categories were presented to participants in a randomized order. In the test phase, ten of these previously seen exemplars from each category were presented to participants along with ten novel chequerboard stimuli from each of the three

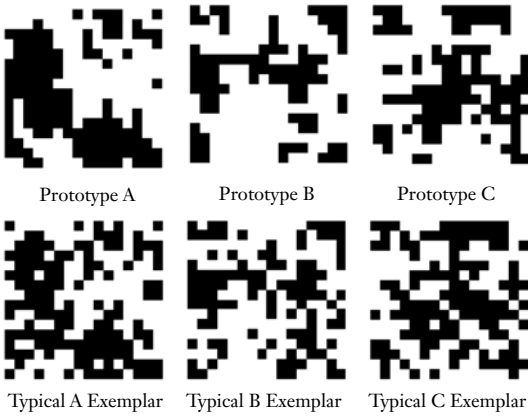

FIGURE 2. Examples of the prototypes (top row) and exemplars (bottom row) from the categories used in the experiments reported in this paper. Please see the text and Civile, Zhao, Ku, Elchlepp, Lavric and McLaren (2014) for more details about the characteristics of our prototype-defined categories of chequerboards.

categories and the three previously unseen category prototypes. The prototype stimuli were presented twice each during test. Participants made category responses to stimuli using the "C", "V" and "B" keys on a keypad.

tDCS

This was delivered by a battery driven constant current stimulator (Neuroconn) using two electrodes covered by 5 cm × 7 cm pieces of pre-dampened synthetic sponge. One electrode montage was used: the first electrode (to which polarity refers) was placed over the left PFC (Fp3) and the reference electrode was placed on the forehead above the right eye. First electrode placement was determined by locating the Cz for each of the subjects (half the distance between the inion and nasion areas) and then moving 7 cm anterior relative to the Cz and 9 cm to its left (see figure 3).

Current was applied 1.5 min before the participants began the categorisation task (whilst listening to instructions) and from then on making 10 min stimulation in total. tDCS was delivered with an intensity of 1.5 mA, and a fade-in and fade-out of 5 sec for the Anodal and Cathodal groups. Sham re-

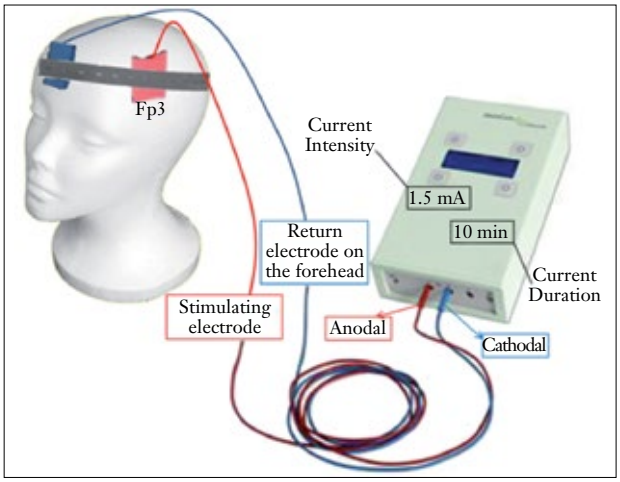

FIGURE 3. The figure illustrates the electrode configuration and the tDCS apparatus used in these experiments.

ceived the same 5-sec fade-in and fade-out, but only 30 sec stimulation between them, which terminated before categorisation commenced. A double blind procedure was used, by having two experimenters, one (primary) who actually ran the participant, and another (secondary) who set up the stimulation according to specifications provided by a third party. The connections to the stimulator were concealed by the secondary experimenter so that neither primary experimenter nor participant could determine the polarity of stimulation. In Experiment 1 we compared Anodal, Cathodal and Sham groups.

Design and procedure

In a between-subjects design the 48 participants were randomly assigned to one of three conditions: anodal stimulation, cathodal stimulation, or sham. Thus, all conditions contained 16 participants.

Once participants had been set up for tDCS stimulation they were informed that they would see different black and white chequerboard stimuli that they had to categorise into category A, B or C, and were shown the three buttons on the keyboard that they were to use ("C", "V" and "B" respectively). After the tDCS stimulator was switched on, the participant then read through three

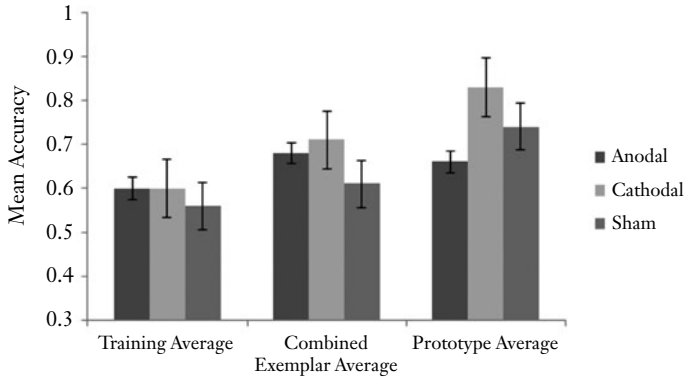

FIGURE 4. Average accuracy for each group (Anodal, Cathodal, Sham) broken down to show overall performance during training and then test performance to exemplars and prototypes. Error bars are SE of the mean.

screens of more detailed instructions about the task, which lasted approximately 1.5 minutes. The training phase then began, which contained 192 novel category stimuli presented in three blocks of 64 randomized trials with self-paced breaks separating each block. After a fixation cross, one stimulus was presented for 3 seconds during which the participant made their category response on the keyboard. The stimuli remained on the screen for the full 3 seconds. Feedback was presented after every trial.

After the participant finished the training phase, the primary experimenter switched off the tDCS stimulator and informed participants that no current was now going through the electrodes. They were then informed that there was a final block to the task, using the same categories as before, but this time with no feedback. This test phase had 66 trials of randomised exemplar and prototype stimuli.

Results

The crucial dependent variable was mean accuracy proportion (out of 1) of category responding during the test phase of the experiment (figure 4).

To examine the prototype effect, the difference between accuracy in responding to exemplar and prototype stimuli during test was investigated.

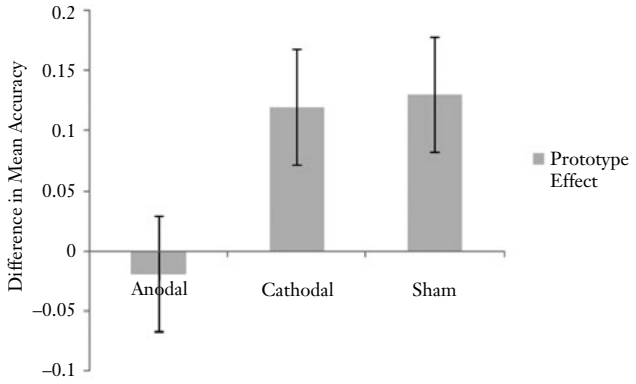

FIGURE 5. The difference in response accuracy between prototypes and the average of responding to exemplars in the test phase of the experiment. Error bars are SE of the mean.

The average accuracy in responding to exemplars during test was calculated for each participant, and the mean accuracy of responding to these exemplars was then subtracted from the accuracy in responding to the prototypes during test (figure 5). This difference was then entered into a univariate analysis as a dependent variable with condition as the fixed factor. The main effect of condition on this measure of the prototype effect approached significance (p = .081). There was no significant difference when comparing cathodal stimulation to Sham. However, when comparing the prototype effect in the anodal stimulation condition to the Sham control group there was a significant difference (p = .03) indicating that the prototype effect was smaller in the Anodal group. Comparing the prototype effect under anodal stimulation to the cathodal stimulation condition there was also a similar significant difference between conditions (p = .043), i.e. a greater prototype effect under cathodal stimulation compared with anodal stimulation. It is the lower accuracy on prototype trials in the anodal condition that seems to be driving these results.

Differences between exemplar and prototype response accuracy were also compared with the null hypothesis of a difference of zero between the two measures. There was no reliable difference found in the Anodal condition, however, Cathodal and Sham conditions both produced significant effects on this test (p < .05) indicating a significant prototype effect for these conditions (see figure 5).

Discussion

Our results are broadly in line with those of Ambrus et al. (2011), in that we have also shown that anodal stimulation at Fp3 leads to a significant reduction in, perhaps even elimination of, the prototype effect. Whilst accuracy scores are significantly higher for the prototype than for exemplars under Cathodal and Sham stimulation, this difference disappears under anodal stimulation and the difference between these differences (i.e. prototype effect for Anodal vs. prototype effect for Cathodal or for Shams) is also significant. Ambrus et al. (2011) also found that anodal stimulation to left DPLFC eliminated a prototype effect that was otherwise significant in Sham controls, though in their case this was accompanied by significantly lower performance to prototypes in the Anodal condition relative to Shams as well, a result that is not significant in our data though the numerical trend is the same. Our results do allow us to extend Ambrus et al.'s conclusions, however, as we have been able to show that cathodal stimulation is not different to sham stimulation with our procedures (Ambrus et al. did not run a left DPLFC cathodal group). Thus, our effect is a selective one, in that only anodal stimulation of left DPLFC eliminated the prototype effect in our experiment.

We will forgo further analysis of this result until we have reported the results of Experiment 2, which also investigates the effects of tDCS to left DLPFC, but this time using a version of our categorisation task that is identical to that used in our earlier perceptual learning experiments (Civile et al., 2014).

Experiment 2

Here we carry out two replications of an experiment that exactly duplicates the categorisation training procedure adopted by McLaren (1997) and also used by Civile et al. (2014). This was done in order that our results could be extrapolated to these perceptual learning experiments, allowing us to predict the consequences of tDCS for perceptual learning in future experiments. In this procedure only two chequerboards are used as base patterns or prototypes (i.e. there are only two categories in play), and exemplars are generated from them as before by adding noise, which simply involves changing a random selection of the squares in the prototype. Participants are then trained

to distinguish between exemplars drawn from these two categories using a trial and error procedure with feedback before being tested for classification accuracy to both category exemplars and their prototypes (which, as in Experiment 1, are never seen in training). Experiment 2a uses this paradigm and contrasts anodal tDCS to Fp3 in the Experimental group with a Sham control. Experiment 2b uses a cathodal stimulation group as the comparison with the Experimental group receiving anodal stimulation. The cathodal control has the advantage that stimulation occurs in exactly the same way as for anodal stimulation (but with reversed polarity). We took this opportunity to see if it would produce similar results to sham stimulation.

Method

Stimuli

These were as before but only two prototype-defined categories were used (A and B in figure 2).

Participants

Experiments 2a and 2b each had 16 undergraduate participants per group and were run in Shanghai, China, at East China Normal University.

tDCS

Stimulation was as in Experiment 1. In Experiment 2a we compared Anodal and Sham groups. In Experiment 2b we compared Anodal and Cathodal groups.

Categorisation task

Participants were asked to categorise chequerboards into two different categories (in this case A and C, see figure 2). Chequerboards were presented one at a time for classification. They were presented for 4 seconds. Partici-

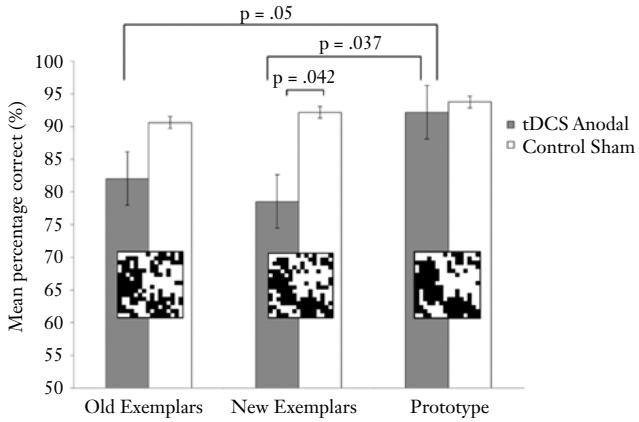

Figure 6. The graph shows mean accuracy during test for old and new exemplars drawn from the trained categories as well as performance on the prototypes for those categories. The chequerboards shown are typical exemplars / the prototype for the A category, but the average is for both categories. Error bars show SE of the mean.

pants had to press either the "x" or the "." key to categorise the stimulus. The experiment moved to the next stimulus only after the 4 seconds had passed. Participants received feedback as to whether their response was correct or not. 128 exemplars were presented, 64 from category A and 64 from category C. In the test phase participants were asked to categorise chequerboards (self-paced) without feedback. They were given one presentation of eight old exemplars from each category (exemplars used in training), eight new exemplars from each category, and two presentations of both category prototypes.

Results. Experiment 2a

Figure 6 gives graphs of mean accuracy for Experiment 2a. A strong prototype effect was obtained under anodal tDCS, but was absent in the Sham group; $p < .05$ for comparisons between the prototype and mean performance on the exemplars in the Anodal condition. The interaction for these effects with group (Anodal vs. Sham) did not, however, reach significance ($p = .15$). There is some evidence that the effect of anodal tDCS was to suppress performance to the exemplars, in that there was a significant differ-

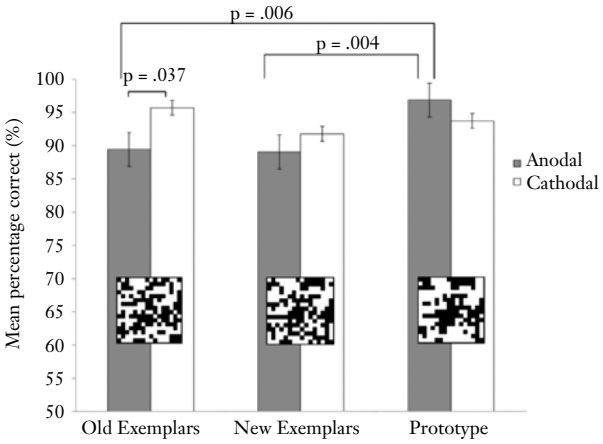

FIGURE 7. The graph gives mean accuracy for old and new exemplars as well as the prototypes during test based on performance on both categories. This time the figure displays typical exemplars / the prototype for the C category.

ence between Anodal and Sham groups for the New exemplars, $p = .042$. Clearly, given our earlier results and those of Ambrus et al. (2011), this set of data came as something of a surprise. Further interpretation of this result will be postponed, however, until we have considered the results of Experiment 2b.

Results. Experiment 2b

Figure 7 gives the graphs for Experiment 2b. Once again a prototype effect was obtained under anodal tDCS, $p = .005$, but not under cathodal tDCS, which gave results very similar to those obtained in the Sham group of Experiment 2a. There was some evidence that the Anodal group prototype effect was significantly stronger than that in the Cathodal group, $p = .078$ for the interaction using the average of the two types of exemplar to compare to the prototype. There is also evidence that anodal tDCS suppresses test performance to exemplars, as there is a significant Group difference, this time for Old exemplars, $p = .037$.

Discussion

Taken together, the results of Experiment 2 suggest that anodal tDCS reduces accuracy on test to exemplars in this type of categorisation task. It leaves performance to prototypes relatively unaffected, however, which leads to the emergence of a prototype effect when we compare performance on the prototype to that on other exemplars. Before accepting these conclusions, however, we acknowledge that there is an obvious issue with these results that makes their interpretation more difficult. Performance in the Sham or Cathodal groups is near ceiling, particularly for the prototypes. This makes it hard to tell whether the absence of any prototype effect in these groups is real — or is due to this ceiling effect. If it is the latter, then it may be that anodal tDCS simply reduces test accuracy below ceiling, allowing a prototype effect that was, in some sense, always there to emerge. Another possibility, however, is that anodal tDCS *selectively* enhances the prototype effect in these experiments, and that it's appearance is not a simple consequence of an overall reduction in performance allowing an effect that was present but masked to become visible. We will focus on this last possibility in what follows, as we have been unable to generate a plausible account of how anodal tDCS could reduce overall performance in the two category case, but selectively reduce performance to prototypes in the three category problem.

On the face of it, the results of Experiment 1 and Experiment 2 appear to be incompatible. In Experiment 2, as we have just seen, we have evidence for anodal tDCS using our electrode montage producing a stronger prototype effect than that shown in our control groups (using either Sham or cathodal stimulation). In Experiment 1 we obtained the converse pattern of results, the prototype effect in the Anodal group was this time significantly weaker (and actually absent) than in either Sham or Cathodal groups. It is true that because of the nature of the problems there are some parametric differences in stimulation between the two experiments. tDCS stimulation will have been active for about half of the training phase in Experiment 1, but the full training phase in Experiment 2. But this, on its own, would seem an unlikely candidate to explain the opposite effects of the two experiments, and in any case the effects of tDCS stimulation are thought to last well beyond the active stimulation period. So how are we to explain this pattern of results?

We believe that the key to understanding this pattern lies first of all with the prototype effect (or lack of it) demonstrated in the control conditions where tDCS can be assumed to not have any significant influence. In the two category problem used for Experiment 2 there was no prototype effect in these control groups. In the three category problem used in Experiment 1 there was a significant prototype effect in both control groups. The stimuli and procedures in both experiments are the same, with the proviso that we used an extra category in Experiment 1, so this difference (no prototype effect vs. prototype effect) can most probably be ascribed to the use of three rather than two categories. This would have the effect of influencing performance levels not only because there are three possible choices instead of two, but also because the amount of generalisation between categories has increased (because now each test stimulus in the three category problem would be receiving generalisation from exemplars of two different categories in addition to members of its own category, rather than from just one).

This extra generalisation between categories would also, somewhat paradoxically, produce a stronger perceptual learning effect for the three category problem than would be the case in the two category problem. The extra generalisation makes the perceptual discrimination between categories more difficult, but the perceptual learning effect addresses this issue, by enabling the representations of the exemplars and prototypes from the three categories to become more distinct, and consequently there is more scope for this effect to manifest in these circumstances We will go into considerable detail on exactly how this might be achieved shortly, but our argument is that this stronger perceptual learning effect in the three category problem is particularly marked between categories, making them more easily distinguishable from one another and this enhances the prototype effect.

Our explanation of the results for the *control conditions* is thus based on a trade-off between generalisation between categories (which on its own reduces classification performance) and enhanced between-category perceptual learning, which we will argue assists classification of prototypes more than exemplars. In the two category problem the former effect dominates, and generalisation between categories is such that it counteracts any advantage that the prototype might have over other exemplars. In the three category problem the balance shifts, and now perceptual learning makes the categories more discriminable and the prototype effect emerges. We will show how this can happen shortly, but note that some explanation for this (relia-

ble) difference between control conditions has to be given, and this is the most plausible account available to us.

Our explanation of the results in the anodal tDCS conditions is that this stimulation abolishes perceptual learning, leaving enhanced generalisation, both between and within categories. The effect of the enhanced generalisation within-category is to strengthen the prototype effect, but the effect of the between-category generalisation will be to reduce it. The first dominates in the two category problem, but the second is the more important factor in the three category problem because the amount of between-category generalisation is doubled. Hence the prototype effect in the two category problem becomes detectable under anodal tDCS (and may be potentiated by a reduction in performance from ceiling — we cannot rule this out); but the prototype effect that was already detectable in the three category problem is reduced and becomes non-significant in the three category case.

The analysis thus far may seem rather ad-hoc and designed to describe rather than explain our data. Note, however, that there has to be some explanation for the otherwise rather counter-intuitive pattern of results obtained across Experiments 1 and 2, and that our explanation of the effects in the control groups follows from an application of the McLaren, Kaye and Mackintosh (1989) model of perceptual learning and categorisation and it's recent variants (McLaren and Mackintosh, 2000; McLaren, Forrest and McLaren, 2012) discussed in our introduction. Our hypothesis is that the modulation of salience based on the error term that forms a vital part of MKM model is disrupted by anodal tDCS so that the model in essence reverts to McClelland and Rumelhart's (1985) model of categorisation inasmuch as perceptual learning or representation development is concerned. This hypothesis is explored in detail in the computational analysis that follows.

Perceptual Learning and Categorisation under tDCS

The top middle panel of figure 8 shows how the salience (activation) of the elements (representations of sets of features) of each category prototype will be affected by experience of exemplars from categories A and C if we adopt the MKM approach to salience modulation via prediction error. Note that all the elements needed to represent all three categories (A, B and C) are shown

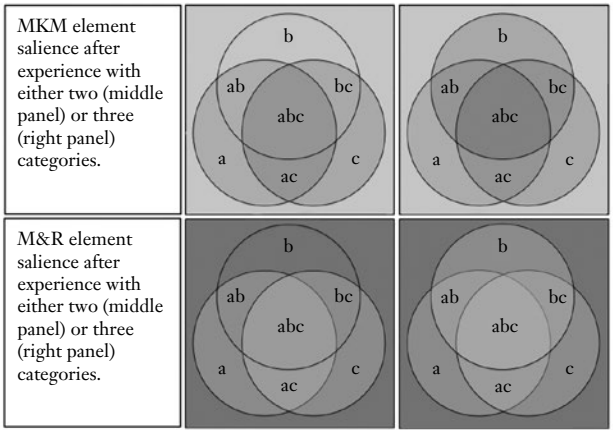

FIGURE 8. Top Panels: This illustrates how modulation driven by prediction error (as in MKM) can be used to influence feature salience. The result, shown in the temperature diagram, is that stimulus features that are more predictable become less active (darker shading), leading to latent inhibition (slower learning as a consequence of pre-exposure). This improves discrimination between members of a prototype-defined category as it relies upon the less predictable features unique to each stimulus. The bottom panels show how removing this modulatory input (as in M&R) reverses this effect, making the prototypical features of the stimuli the most salient (lighter shading). The two category case (centre) where exposure is only to A and C categories, and the three category case (right) are illustrated in terms of the prototypes for each category, and are labelled to show the differential effect on the elements that make up each category prototype.

for completeness, but that exposure is only given to two of them, A and C, for this example. Those elements that are more predictable and are more often encountered will be those with lower salience (darker shading). Thus, the elements shared by the A and C prototypes (abc and ac) are less salient than a or ab elements (only present in A). Given that exemplars from the B category are not pre-exposed in this example the b elements can only occur by virtue of the random noise added to construct exemplars from the prototypes. The right panel shows how the modulation of salience across elements changes when all three categories are experienced. In particular, the shared prototypical elements, abc, become even less salient. The effect is that discrimination between

Table 1

Stimulus	*Elements*	*a*	*ab*	*ac*	*abc*	*bc*	*b*	*c*	*n*
A prototype	%	25	25	25	25	0	0	0	0
A exemplar	%	20	20	20	20	5	5	5	5
C prototype	%	0	0	25	25	25	0	25	0
C exemplar	%	5	5	20	20	20	5	20	5

Note. This shows the percentage of features of different types (elements) present in each of four different stimuli drawn from two different categories. Each element label refers to figure 7 (middle and rightmost panels) and denotes features that are present in one or more of the three possible categories. The table also makes clear the extent of feature overlap between any two stimuli once it is born in mind that exemplars are generated from the prototypes by randomly changing the elements of the prototype, and that 25% is the maximum allocation of elements of any one type.

the three category prototypes is actually better than when only two categories were trained because perceptual learning is more effective.

The bottom panels of figure 8 show what happens when the salience modulation mechanism in MKM is removed. The salience (activation) of elements representing the stimulus features now reverts to that in McClelland and Rumelhart's (1985) model of categorisation, with units receiving more internal input having higher (rather than lower as in MKM) activations. In effect, this gives the common elements an advantage that can be seen in both lower panels. They become increasingly salient, and this leads to very strong between- and within-category generalisation. Table 1 gives the relative proportions of the different elements making up each stimulus for the average A exemplar and C exemplar as well as the A and C prototypes using a simple model that, as a first approximation, equates each square in a chequerboard with a feature. By combining this information with the expected salience of these elements shown in figure 8, it is possible to get a sense of how much one stimulus will generalise to another as a result of categorisation training.

We can see immediately that the MKM model predicts that exemplars will contain novel (noise) elements that are of relatively high salience, and that the prototypical elements will be more numerous, but less salient. The prototypes are exclusively composed of relatively low salience prototypical elements and do not overlap as much as exemplars drawn from the two categories. The consequence of this is that generalisation from, say, the trained A exemplars to C exemplars will be somewhat greater than to the C prototype. This effect is symmetrical (the C exemplars generalise to the A exem-

TABLE 2

Stimulus	*2 Categories MKM*	*2 Categories M&R*	*3 Categories MKM*	*3 Categories M&R*
A prototype	.596	1.00	.395	1.00
A exemplar	.609	.754	.526	.761
C prototype	.340	.654	.173	.672
C exemplar	.430	.558	.422	.518

Note. This shows the expected generalisation (minimum=0, maximum=1) between each of the four stimuli in the table and a typical trained exemplar drawn from Category A. Generalisation is calculated using either the MKM or M&R salience for the elements comprising the stimulus. Note that the generalisation from A exemplars to the C prototype will be the same as that expected for the C exemplars to the A prototype, and so can be used in conjunction with the figures for the A prototype to calculate the probability of labelling prototype A as a member of the A category.

plars to the same extent), and so the chance of mistakenly calling an A exemplar a member of the C category will be somewhat greater than that of calling the A prototype a member of the C category.

Table 2 gives the calculated expected generalisation (based on figure 8 and Table 1) to/from trained A exemplars to each of the four stimulus types considered in our earlier table, and it confirms our analysis. If we begin by looking at the 2 Categories MKM column of the table, it shows (perhaps rather surprisingly) that the generalisation from one of the trained A exemplars to this typical A exemplar (.609) will be greater than that stimulus' generalisation from (or to) the A prototype (.596), but the difference between these values is not large. We can estimate the generalisation that occurs on average from the C category exemplars and the C prototype to/from this A exemplar by looking at the C prototype and C exemplar rows of the table. These give generalisation from an A exemplar to these stimuli, but by symmetry they give us the values we will require for our calculations. Thus, the generalisation from C exemplars to an A exemplar (.430) will be considerably greater than the generalisation from the C prototype to A exemplars (.340), which is also the value for generalisation from C exemplars to the A prototype. The result is a larger difference in generalisation for the A prototype to the A category exemplars compared to the C category exemplars (.596 – .340 = .256) than for the A exemplars to the same stimuli (.609 – .430 = .179). In other words, it predicts a prototype advantage, but does it predict a detectable prototype effect?

To answer this question we need to convert generalisation into choice. The models themselves do not stipulate the requisite decision mechanisms to function as stand-alone classifiers. Hence we used a minimalistic approach to converting generalisation into choice behaviour that was simply designed to demonstrate that the MKM model could produce the correct pattern for the two and three category problems in the control groups, and that this would then change appropriately when error modulation of salience was disrupted. We employed a standard form of Luce's choice rule, using the exponential of the generalisation coefficient as our measure of category membership.

$$P(A) = \frac{e^{ka}}{e^{ka} + e^{kc}} \qquad 1$$

Where P(A) is the probability of classifying a stimulus as a member of category A, a is the summed generalisation to that stimulus from trained A exemplars, c is the summed generalisation from trained C exemplars, and k is a constant that captures the weight given to generalisation in a given task. We then needed to find k for our model. For the 2 Categories MKM coefficients we simply chose k so that it gave a ballpark fit to the accuracy data for the exemplars in our experiments (and we used this procedure for the other data as well). We adopted a value of 11, which resulted in P(A) for the prototypes being 0.94, and P(A) for exemplars being 0.88. These are a reasonable fit to the actual values across the two experiments, which are 0.94 and 0.925 respectively, though clearly the model value for the exemplars is a little low.

One point to make here about this very simple model is that we are simply assuming that each square in a chequerboard is a feature. This may be a useful approximation to reality for our purposes, but it completely fails to capture the fact that the prototypes (which were constrained to have regions of nearly all black or nearly all white) looked distinctly different to the exemplars, even those from their own category, which were necessarily less "blocky" in appearance because of the random noise used to generate them (see figure 2). This would act to reduce the magnitude of any prototype effect in these experiments, and so our model is necessarily overestimating the size of the prototype effect actually obtained. Even given this, however, we can see that a prototype effect might be hard to detect for the two category case under our control conditions.

We can now look at the expected generalisation for the three category problem. This is shown in the 3 Categories MKM column, and gives a dif-

ference of .222 (= .395 – .173) for the prototype and .104 (= .526 – .422) for the exemplar. Clearly this is a larger disparity between prototype and exemplar generalisation (.118) than we had for the two category case (.077 where the values for the prototype and exemplars were .256 and .179), and as such could lead to a stronger prototype effect. We took k for the three category task to not necessarily be the same as in the two category task, and arrived at a value of 10. Clearly training on three categories rather than two might, in itself, affect the weight placed on the measure provided by generalisation (not least because as the number of categories increases so does total generalisation between them), but note that using the same value for k as in the two category case (i.e. 11) leads to essentially the same pattern of results with this simple model of choice. The choice equation now becomes:

$$P(A) = \frac{e^{ka}}{e^{ka} + e^{kb} + e^{kc}} \qquad 2$$

This resulted in P(A) for the prototypes being 0.82 (which is somewhat too high), and P(A) for exemplars being 0.59 (which is too low), but represents a reasonable fit to the data and clearly makes the point that the prototype effect for the three category problem is predicted to be much greater than for the two category problem (a difference of 0.82 – 0.59 = 0.23 compared to a difference of 0.06 in the two category problem). It is no surprise on this analysis, then, that the prototype effect might be detectable in our controls for the three category, but not the two category problem.

If we now consider the effect of turning off error-based modulation of salience to give something like the representation development that would be seen using the M&R model, then a quite different pattern emerges. First of all, generalisation increases a great deal — as can be seen by looking in the two M&R columns of Table 3. This is exactly as would be expected given that perceptual learning (which has effectively been switched off) has the opposite effect to generalisation. The increased generalisation for the two category problem gives difference scores of 1 – .654 = .346 for the prototype and .754 – .558 = .196 for exemplars. The difference score for the prototype has improved relative to the .256 difference obtained using MKM, whereas the score for the exemplars has stayed about the same (it was .179). The prediction, then, is that the prototype effect should be enhanced by anodal tDCS in the two category case, as the disparity between prototype and exemplar difference scores is now .150 instead of the original .077. Translating the

generalisation scores into choice probabilities requires that we make a new estimate of k here, as clearly tDCS could quite possibly have affected the weight placed on our measure of category membership in ways not captured by our model. A value of k = 8 gives us P(A) for the prototype as 0.94 and P(A) for exemplars as 0.82 in the two category problem, which is a good fit to our data and suggests that the size of the predicted effect has doubled. If we instead consider what happens for the three category problem then a different effect emerges. The original disparity between the generalisation differences for MKM was .118 (.222 – .104), but once we turn off error-based modulation it becomes .085 (.328 – .243). Clearly both generalisation scores have increased, but the increase has been greater for the exemplars and so the difference is smaller, and smaller still relative to the scores contributing to that difference. Translating these scores into choice probabilities we used a value for k of 5 to try and fit our data as best we could, which results in choice probabilities of 0.72 (too high) for the prototype and 0.63 (too low) for the exemplars. Clearly, this simple model of choice had considerable difficulty in fitting our data. But this exercise makes the important point that once again the changes in generalisation — which are all we are confident of in this modelling exercise (and even here we have caveats about the similarity of our prototypes to the exemplars) — do translate into changes in choice probability which fit the interaction in our data. In this case the predicted prototype effect, which was 23%, has now decreased to 9% indicating that it should become considerably more difficult to detect.

One point that may strike the reader about our analysis is that this final prototype effect for the three category problem under tDCS (an effect of 9%) is not so different to the effect of 12% predicted for the two category problem under tDCS which we wish to claim is detectable. The important points to make here are that first, the two effects occur at different levels of choice probability. A 12% difference when choice is in the 80%-90% range will have a lower variability associated with it than a difference of 9% when choice probabilities are 60%-70%. Thus one may be detectable where the other is not. Second, we have tried to emphasize that it is the *change* in effect from control stimulation to experimental (anodal) stimulation that is the real prediction of interest here. We cannot (and do not wish to) lay claim to possessing a model that fits (in the statistical sense) our data, but we can claim that the changes in generalisation that occur in our model as a result of shifting from two category to three category problems, and as a result of switch-

ing off perceptual learning, accurately capture the pattern in our data. And this suggests a particular interpretation for the effects of anodal tDCS stimulation of DLPFC.

The real test of our position, of course, would be to look directly at the effects of anodal tDCS stimulation on perceptual learning. We are now able to unequivocally predict that this stimulation should disrupt perceptual learning and possibly even reverse it. We will now briefly consider a set of experiments that addresses this issue, using the same set of chequerboard stimuli and the design employed by Civile et al. (2014) to look at perceptual learning in the context of inversion effects.

Perceptual Learning

We have already noted that perceptual learning affects the way we see the world and the objects in it, and that pre-exposure to stimuli enhances our ability to discriminate among or between them or other similar stimuli. In the lab, one of the most striking consequences of perceptual learning is the face inversion effect: upright faces are better recognised than inverted faces. This inversion effect is at least partly due to our extensive experience with faces, as exposure to artificial stimulus sets that have a structure akin to that possessed by faces leads to phenomena similar to those observed in face recognition, including inversion effects (McLaren, 1997; Gauthier & Tarr, 1997). For example, exposure to a set of prototype-defined chequerboards results in an inversion effect for exemplars from a familiar category but not for exemplars from a novel (not pre-exposed) category (McLaren, 1997, Civile et al., 2014). As we have already argued, this advantage for upright exemplars can be explained by associative models of perceptual learning that rely on differential latent inhibition of common elements. Exposure to exemplars from the familiar category leads to latent inhibition of the prototypical elements for that category (figure 1, top half). When an exemplar drawn from that category is encountered, the elements that it shares with the prototype will be latently inhibited (making them less salient), whereas the elements that are unique to that exemplar will not suffer greatly from latent inhibition (making them more salient). This will enhance discrimination between exemplars drawn from the familiar category (i.e. perceptual learning). Our associative model can explain a range of perceptual learning phenomena, including the

inversion effect, as the latent inhibition mechanism only applies to what has been experienced, and participants have not experienced inverted exemplars during the earlier familiarization phase. Figure 1 (bottom half) also shows that losing the modulatory component producing differential latent inhibition should result in a loss and perhaps even a reversal of within-category perceptual learning.

Our plan, then, was to run what was essentially a replication of Civile et al. (2014), but to apply anodal tDCS using our current electrode montage during the first, categorisation phase. This should disrupt any perceptual learning and increase generalisation between exemplars. Because perceptual learning is responsible for the inversion effect for exemplars drawn from a familiar category (i.e. one that has been trained) that we reliably see with this procedure, anodal tDCS should reduce (perhaps even reverse) this effect. The detailed results of these experiments will be reported elsewhere (Civile, Verbruggen, McLaren, Zhao, Ku, & McLaren, in preparation) so we will only summarise them here. We used the same 16 × 16 chequerboards used in the earlier experiments with the addition of one extra category, D. Our experimental groups used anodal stimulation. The control groups used Sham or cathodal stimulation. Participants classified exemplars from two prototype-defined chequerboard categories during tDCS (categorisation stage). They then studied exemplars drawn both from one of the now familiar categories and from another novel category in either upright or inverted orientations (study stage). Finally, in the recognition task (test stage) they had to classify chequerboards as either "old" (seen in the study phase) or "new" (not seen). Their accuracy scores were then converted into d' measures for use in our analyses.

In figure 9 we give the combined Anodal stimulation vs. Control results for recognition in this final phase. As predicted by extrapolation from the Civile et al. (2014) experiments and the results considered earlier, in the Control conditions we observed an inversion effect for familiar-category exemplars (Upright better than Inverted, $p = .013$) but not for novel-category exemplars. The perceptual learning effect was also reflected in the performance on upright exemplars taken from the familiar category being better ($p = .050$) than that on the matched exemplars (matched across participants) taken from the novel category. But under anodal stimulation the pattern was quite different. There was no inversion effect, and the effect that was there was significantly different to that in controls ($p = .045$). Now the performance on upright exemplars taken from the familiar category was significantly worse ($p = .005$)

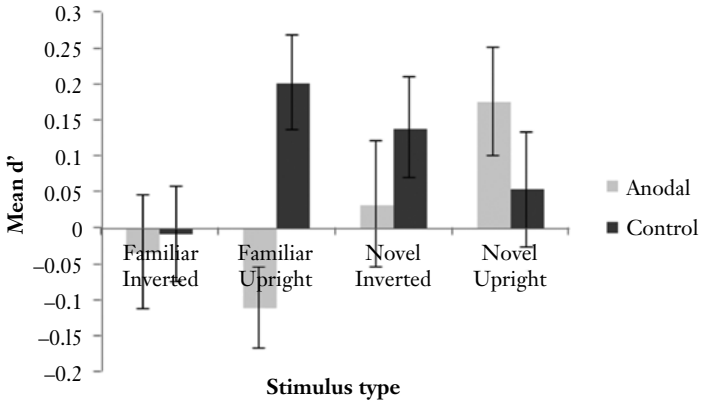

FIGURE 9. Combined results of perceptual learning experiments. Lighter bars are for anodal stimulation and darker bars for control stimulation. The y-axis gives d' scores for the old/new recognition task (higher = better, 0 = chance), and the four different stimulus conditions are shown on the x-axis.

than that on the matched exemplars taken from the novel category. In fact, if we compare performance on the upright exemplars taken from the familiar category in both conditions, the difference is also highly significant ($p = .005$), and in favour of the controls. Clearly the effects of familiarisation with the category have radically altered under anodal stimulation.

The most reasonable interpretation of these results is that Anodal tDCS has a selective effect on performance to upright exemplars drawn from the familiar category. We would argue that our data are consistent with anodal tDCS stimulation eliminating a modulatory input based on prediction error, leading to a loss of perceptual learning. The resultant system is then adequately described by simple delta rule algorithms of the type found in the M&R model of categorisation, and as such has a particular problem in dealing with familiar prototype-defined categories as a consequence of the increased generalisation between their exemplars. This leads to the poor performance on the upright exemplars drawn from the familiar category under anodal tDCS, compared to the otherwise superior performance exhibited to these exemplars under control conditions as predicted by MKM.

Conclusions

In Experiment 1 we were able to demonstrate that anodal tDCS to left DLPFC does indeed reduce the prototype effect that might otherwise be obtained after learning to categorise. This confirms the result of Ambrus et al. (2011), and suggests that their result was not simply a matter of anodal tDCS reducing learning *per se*. We carried out Experiment 2 in order to set the stage for our subsequent investigation of perceptual learning and to confirm the results of Experiment 1. Our results were, on the face of it, anything but confirmation of Experiment 1, in that far from reducing the prototype effect, anodal tDCS enhanced it. In fact, it produced a significant effect where under control conditions none had been detectable.

This initially surprising and contradictory result proved to be susceptible to a detailed analysis in terms of changes in generalisation brought about by 1) changing the number of categories from three to two, and 2) using either anodal or control stimulation. The analysis relied on the assumption that the effect of anodal tDCS to left DLPFC was to disrupt modulation of the salience of stimulus representations based on error such as to transform a system for categorisation that under control conditions could be described by MKM, to one better thought of in terms of M&R. This effect interacted with the increased generalisation between categories that occurred in the three category problem relative to the two category version, and so explained the different effects of anodal tDCS on the prototype effect in the different experiments. Our analysis is model-driven, and admittedly post-hoc, but it did make the prediction that perceptual learning due to pre-exposure during categorisation training should be eliminated, or even reversed, by anodal stimulation.

This prediction was fully borne out by the results of the final set of experiments reported here. Our control conditions showed our usual inversion effect in this analogue of the face recognition paradigm using chequerboards, but the inversion effect was not present under anodal tDCS. More importantly, whilst familiarisation with a category improved performance on upright exemplars drawn from that category, in that they were better discriminated in the old/new test than those drawn from a novel category, this effect was reversed under anodal stimulation. Finally, performance on the upright exemplars from the familiar category was significantly and selectively worse under anodal tDCS than in the control conditions, an effect entirely consistent with

our hypothesis that anodal stimulation "turns off" perceptual learning and leaves participants with greatly increased generalisation.

Final Thoughts

The McLaren, Kaye and Mackintosh theory of latent inhibition and perceptual learning could, up until now, be seen as an abstract connectionist model of representation development that provided a good account of a fairly limited domain of animal and human behaviour. But the intention on the original author's part was always to apply it more widely, and that is a challenge that we have taken up with our recent research into categorisation and perceptual learning. We now have some hints about the neural mechanisms underlying perceptual learning, and they fit very well within the framework provided by that theory. This serves to remind us that Nick Mackintosh's vision in extrapolating from sophisticated behavioural experiments to detailed theoretical mechanisms was quite extraordinary, and his theoretical insights into perceptual learning in humans are as relevant today as they were over twenty-five years ago.

References

Ambrus, G. G., Zimmer, M., Kincses, Z. T., Harza, I., Kovacs, G., Paulus, W., & Antal, A. (2011). The enhancement of cortical excitability over the DLPFC before and during training impairs categorisation in the prototype distortion task. *Neuropsychologia*, *49*, 1974-1980.

Civile, C., Zhao, D., Ku, Y., Elchlepp, H., Lavric, A., & McLaren, I. P. L. (2014). Perceptual learning and inversion effects: Recognition of prototype-defined familiar chequerboards. *Journal of Experimental Psychology: Animal Behavior Processes*, *40*, 144-61.

Civile, C., Verbruggen, F., McLaren, R., Zhao, D., Ku, Y., & McLaren, I. P. L. (in preparation). Switching off perceptual learning: tDCS to left DLPFC eliminates perceptual learning in humans.

Fregni, F., Boggio, P. S., Nitsche, M. A., Bermpohl, F., Antal, A., & Feredoes, E. (2005). Anodal transcranial direct current stimulation of prefrontal cortex enhances working memory. *Experimental Brain Research*, *166*, 23-30.

Gauthier, I., & Tarr, M. G. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *12*, 1673-1682.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalisation and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 418-39.

Kincses, T. Z., Antal, A., Nitsche, M. A., Bártfai, O., & Paulus, W. (2003). Facilitation of probabilistic classification learning by transcranial direct current stimulation of the prefrontal cortex in the human. *Neuropsychologia*, *42*, 113-117.

Lupyan, G., Mirman, D., Hamilton, R., & Thompson-Schill, S. L. (2012) Categorisation is modulated by transcranial direct current stimulation over left prefrontal cortex. *Cognition*, *124*, 36-49.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*, 159-197.

McLaren, I. P. L. (1997). Categorisation and perceptual learning: An analogue of the face inversion effect. *The Quarterly Journal of Experimental Psychology*, *50A*, 257-273.

McLaren, I. P. L., Leevers, H. L., & Mackintosh, N. J. (1994). Recognition, categorisation and perceptual learning. In C. Umilta, & M. Moscovitch (Eds.), *Attention & Performance XV*.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: Latent inhibition and perceptual learning. *Animal Learning and Behavior*, *38*, 211-246.

McLaren, I. P. L., Forrest, C. L., & McLaren, R. P. (2012). Elemental representation and configural mappings: Combining elemental and configural theories of associative learning. *Learning and Behavior*, *40*, 320-333.

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). *An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition*. Oxford University Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 353-363.

Reinhart, R. M. G., & Woodman, G. F. (2014). Causal control of medial-frontal cortex governs electrophysiological and behavioral indices of performance monitoring and learning. *Journal of Neuroscience*, *34*, 4214-27.

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.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141-145.

Considering the Challenge of Mackintosh 2009: (Un)self-supervised Perceptual Learning?

Dominic Michael Dwyer
Cardiff University, UK

Abstract. One critical distinction in the analysis of perceptual learning is between supervised and unsupervised mechanisms. While many experiments with non-human animals can be truly said to involve unsupervised exposure to stimuli, and have thus contributed a great deal to the investigation of unsupervised perceptual learning (in particular through manipulating the schedule by which stimuli are exposed), the same may not be true of apparently similar human studies. Mackintosh (2009) noted that many experiments with humans explicitly or implicitly encouraged participants to look for differences between stimuli during "simple" non-reinforced exposure — something that could support processes of self-supervision or self-reinforcement when these differences were discovered. Subsequent studies provide direct evidence for self-supervision effects in human perceptual learning demonstrating the reality of the issue. However, a review of fMRI-based studies suggests that, even in the presence of possible self-supervision, additional unsupervised mechanisms might still contribute to exposure schedule effects in perceptual learning. Moreover, new data presented here suggests that even when instructions are used to minimise the potential for self-supervision, exposure schedule effects on perceptual learning remain. Thus the challenge of Mackintosh (2009) has been met: unsupervised learning does contribute to exposure schedule effects in human perceptual learning (albeit that the mechanisms for these effects remain to be determined).

Introduction and the Challenge of Self-supervision

It is both appropriate and somewhat ironic that I am taking perceptual learning as the general topic for my contribution to this memorial volume for my PhD supervisor Nick Mackintosh. I arrived in Cambridge to begin my PhD studies at a time when perceptual learning was a central concern for the lab

(e.g., Espinet, Iraola, Bennett, & Mackintosh, 1995; Mackintosh, Kaye, & Bennett, 1991; McLaren, Kaye, & Mackintosh, 1989; Trobalon, Chamizo, & Mackintosh, 1992) and it had been assumed that I would join this ongoing theme of research. In my naïve enthusiasm, I had other ideas, in particular a burning desire to address the question of when and how rodents might learn about the representations of absent cues. Perhaps encouraged by the potential involvement of learning about the retrieved representations of cues in "the Espinet effect", Nick humoured my interests and proved to be as wise, encouraging and influential a supervisor for what became my PhD research as he would have been for research on his preferred topic (Dwyer, 1999, 2000, 2001, 2003; Dwyer, Mackintosh, & Boakes, 1998). But despite his indulgence of my choice of PhD research, he also encouraged me to contribute to the lab's ongoing work on perceptual learning (Dwyer, Bennett, & Mackintosh, 2001; Dwyer & Mackintosh, 2002a, 2002b), and it is instructive as to Nick's foresight and guidance that perceptual learning has proved to be a particularly enduring research interest in the years since I left Cambridge (e.g., Dwyer, Mundy, & Honey, 2011; Dwyer, Mundy, Vladeanu, & Honey, 2009; Jones, Dwyer, & Lewis, 2015; Mundy, Dwyer, & Honey, 2006).

One well established cliché is to begin a discussion of perceptual learning either by quoting Gibson's classic definition as "any relatively permanent and consistent change in the perception of a stimulus array, following practice or experience with this array" (Gibson, 1963, p. 29) or by citing one of Gibson's classic demonstrations that simple, non-reinforced, exposure to stimuli improves subsequent discrimination between them (Gibson & Walk, 1956; Gibson, Walk, Pick, & Tighe, 1958). Perhaps the benefit of such a beginning is that it appears to do little to constrain the topic. However, with his habitual eye for detail, Mackintosh's own consideration of Gibson's apparent generalities revealed a particular aspect of concern (Mackintosh, 2009): namely that for perceptual learning to be distinct from the usual topics of associative learning through the process of reinforcement, it must occur through genuinely unsupervised exposure. This is clearly the case in Gibson's classic experiments, where the stimuli were simply placed in the animals' home cages. It is also the case in many other animal-based studies where, for example, the stimuli were flavour mixtures presented for free consumption to the animals with no consequences (e.g., Blair & Hall, 2003; Hall, Blair, & Artigas, 2006; Mackintosh et al., 1991; Symonds & Hall, 1995). However, even in the absence of explicit feedback or reinforcement directed to the different stimuli,

in many human studies there are opportunities for what Mackintosh termed "self-supervised" learning. That is, either the explicit instructions or the general experimental situation encouraged human participants to look for differences between the stimuli they were being presented — and that when these differences are detected they will be reinforced by virtue of achieving the goal that was (more or less explicitly) set for them. In his 2009 paper, Mackintosh took this "problem" with human research as a key justification for the interest in work with non-human subjects, where the issue of self-supervised learning did not arise. However, for human-based research the challenge remains: does perceptual learning exist in humans in the absence of self-supervision, and if so, how do the mechanisms underpinning the process relate to those identified in other animals? But before considering the problem of self-supervision in detail, it is worth outlining the central features of the analysis of perceptual learning from the associative perspective to which Nick Mackintosh contributed so much.

Notes on Terminology and the Schedule of Exposure to Stimuli

It is common to consider difficult to discriminate stimuli as overlapping collections of elements, where the difficulty of discrimination comes from the fact that the stimuli share a number of common elements (making the stimuli similar) alongside some that are unique (making them at least partially distinct). Such stimuli might be described as AX and BX (where A and B refer to their unique elements and X to the elements they have in common).[1] Often, the distinction between common and unique elements reflects that fact that the stimuli are explicitly constructed as compounds of simpler features: such as salt-lemon and sucrose-lemon flavour compounds (e.g., Mackintosh et al., 1991) or checkerboards constructed by placing one of a number of distinct features on a common background image (e.g., Lavis & Mitchell, 2006).

1. Perhaps associative theorists are overly fond of these "algebraic" descriptions as non-specialists can find the lists of As, Bs, Xs, and Ys impenetrable. Regardless, I will use this abstract terminology here to exemplify how convenient and concise it can be, while attempting to avoid further contributions to "the barbarous terminology" that comprises "one of the most repellent features of the study of conditioning" (Mackintosh, 1983, p. 19).

Associative analysis of perceptual learning centres on the effects that exposure has on the representation of these unique and common elements and the relationships between them. Most generally, exposure could improve discrimination between stimuli if it emphasised selective responding to the unique elements over responding to the common elements (Gibson, 1969).

One of the simplest possible explanations for perceptual learning is that the discriminability of stimuli is a direct function of the frequency with which the to-be-discriminated stimuli have been encountered (i.e., perceptual learning is a simple product of familiarity, e.g., Gaffan, 1996; Hall, 1991). While the amount of exposure is clearly an important factor in any form of learning, it cannot be a complete explanation. Indeed, one of the first contributions by Mackintosh to perceptual learning research was the demonstration (in rats) that the discrimination between AX and BX was improved by exposure to X alone (Mackintosh, Kaye, & Bennett, 1991). Here, exposure to the common element alone (i.e. X) does not affect the familiarity of the unique features (i.e. A and B) upon which the ability to discriminate AX and BX must be based, and so familiarity per se cannot explain the exposure-dependent improvement in discrimination.

Further evidence against the idea that familiarity alone influences perceptual learning comes from the analysis of studies in which the schedule of exposure was manipulated while the total amount of exposure to the relevant stimuli (and hence their overall familiarity) was held constant. For example, exposure on an intermixed schedule (i.e. AX, BX, AX, BX,...) can be matched to the total amount exposure on a blocked schedule (i.e. AX, AX,...BX, BX,...), and yet have significant impacts on the degree to which exposure produces a perceptual learning effect. The first demonstration of such an effect was in chicks, where intermixed exposure to two stimuli resulted in better subsequent discrimination between them than did the equivalent amount of exposure given in separate blocks (Honey, Bateson, & Horn, 1994). This advantage for intermixed over blocked exposure schedules has proved to be highly reliable in both rats (e.g., Bennett & Mackintosh, 1999; Symonds & Hall, 1995) and humans (e.g., Dwyer, Hodder, & Honey, 2004), and cannot be reduced simply to differences in the frequency of exposure (e.g., Mitchell, Nash, & Hall, 2008). The generality of this intermixed/blocked effect across species and stimuli suggests that the manner in which stimuli are exposed is critically important for perceptual learning over and above the simple amount of exposure. Moreover, the analysis of the potential mechanisms underpinning this

schedule effect has been the main focus of research on perceptual learning within an associative tradition. Critically, this analysis has typically assumed that perceptual learning is an unsupervised process in both humans and other animals. Therefore, the root of the challenge offered by Mackintosh's identification of self-supervision is whether exposure schedule effects are present in human experiments where self-supervision does not play a role.

Establishing the Scope of the Self-supervision Problem

Given the deeply entrenched empiricism apparent in his seminal works on associative learning (Mackintosh, 1974, 1983), it is interesting that in his 2009 paper the potential problem of self-supervision was identified not through any direct experimental test, but rather by a hypothetical consideration of the potential impact of task instructions on the way in which human participants might approach a perceptual learning study. However, empirical verification was not long delayed: at least in studies where the to-be-discriminated stimuli comprise unique features superimposed on a common checkerboard background, there is now good evidence that participants' performance is driven by explicitly searching for and attending to the location at which the unique elements appeared during initial exposure (Jones & Dwyer, 2013; Wang, Lavis, Hall, & Mitchell, 2012). Moreover, with the same type of visual stimuli, there is also recent evidence that perceptual learning effects which emerge when participants are instructed to look for differences, which should promote self-supervision, during the pre-exposure stage are not seen when the stimuli are simply presented without instructions or are part of a masking task, which should not (Navarro, Arriola, & Alonso, 2016). Thus the problem posted by Mackintosh (2009) is real. That said, self-supervision is not a problem for all investigations of perceptual learning in humans (something also noted by Mackintosh, 2009): the numerous demonstrations of perceptual learning effects using stimuli that are below perceptual threshold and/or presented as distractors or as "task-irrelevant" stimuli means that neither explicit feedback, nor implicit self-supervision, is a necessary requirement for perceptual learning (e.g., Goldstone, 1994; Tsushima & Watanabe, 2009; Watanabe & Sasaki, 2015). Thus the existence of perceptual learning without either explicit reinforcement or self-supervision is also real. As self-su-

pervision can contribute to perceptual learning, but that perceptual learning can occur without it, it is important to identify where the potential contribution of self-supervision will be most acute in terms of its influence on the analysis of the mechanisms underpinning perceptual learning.

As I have discussed elsewhere (Dwyer & Mundy, 2016) there are two broad approaches to perceptual learning research, one based in a psychophysical tradition, and the other based within an associative learning tradition. As noted above, the most unique contribution of the associative stream of research has been the examination of the ways in which the schedule of exposure influences the process of perceptual learning. Because the effects of expose schedule might be mediated by many different mechanisms, including those supported by self-supervision, the problem of self-supervision is most acute for the associative tradition and it is the research within this tradition on which I will focus.

Manipulating Self-supervision through Instructions

Although the potential problem of self-supervision was so clearly expressed by Mackintosh (2009; see also, Mitchell, Kadib, Nash, Lavis, & Hall, 2008), in the following years it has received almost no direct experimental attention. Indeed, the only published study of which I am aware that has attempted to manipulate the propensity for human participants to deliberately search for differences between stimuli during otherwise unsupervised exposure is the previously mentioned work by Navarro et al. (2016). The key exposure schedule effect — namely improved performance after intermixed as opposed to blocked exposure — was only seen when the participants were explicitly instructed to "indicate whether the stimuli of each pair are the same or different" in the exposure phase (Experiment 1). There was no effect of exposure schedule when subjects were instructed to "simply observe" the stimuli (Experiment 2), or to "count the number of dark-blue splotches" in the checkerboards (Experiment 3). That is, in Experiment 1 the participants were explicitly instructed to look for differences and could be expected to self-supervise their performance, while in Experiments 2 and 3 there was no requirement to look for differences which should act to reduce any tendency to self-supervise. These results are certainly consistent with the idea that the inter-

mixed vs blocked exposure schedule effect is enhanced when there is encouragement and opportunity for the participants to self-supervise.[2]

However, there are several reasons to question whether these results conclusively establish the idea that exposure schedule effects can *only* be seen following self-supervision. Firstly, the experiments were performed using stimuli which comprised a randomly patterned coloured checkerboard background as the common element (X), with the unique elements comprising a cross of red squares that was superimposed on either the lower-left (A) or upper-right (B) quadrant of the background. There is good evidence that discrimination with these sort of stimuli is dominated by the deliberate identification of the location at which the unique elements occur (Jones & Dwyer, 2013; Wang et al., 2012), which implies that these stimuli will be particularly susceptible to self-supervision effects. But this raises the question of whether stimuli that are not so driven by effortful search mechanisms would display the same pattern of effects. Secondly, the change in instructions in Experiment 2 did not only remove the intermixed vs. blocked schedule effect, it also removed any beneficial effects of exposure at all. Thus, although there was evidence from eye-tracking data that participants were looking at the stimuli to some degree, it is possible that the instructions reduced engagement with the task to the extent that the exposure phase was simply ineffective. Finally, the instructions in Experiment 3 required participants to focus on the common background elements. While exposure to the common background checkerboard alone can certainly improve discrimination (Wang & Mitchell, 2011; see also, Mackintosh et al., 1991; Mundy, Honey, & Dwyer, 2007), the presence of the background is entirely equivalent for intermixed and blocked schedules. Thus, to the extent that the instructions in Experiment 3 promoted focusing on the common background alone during the exposure phase, they might have acted to reduce the effective difference between intermixed and blocked exposure. While none of these issues questions the fact that Navarro et al. (2016) only saw perceptual learning effects when self-supervision was

2. Recio, Iliescu, Mingorance, Bergés, and Hall (2015) report conceptually similar results using the same type of checkerboard stimuli: intermixed exposure produced superior discrimination than blocked exposure only in participants explicitly instructed to look for differences between stimuli in the exposure phase; participants instructed either to simply look at the stimuli, or instructed to press a key as soon as the stimuli appeared on screen, did not show this effect.

encouraged, they do raise the possibility that this pattern of results might be very specific to the procedures used in that study and not be indicative of a general influence for self-supervision in human perceptual learning.

New Data on (the lack of) Instruction Effects

To illustrate the potential for the impact of self-supervision to vary across different perceptual learning tasks/stimuli I will now present some previously unpublished work. This experiment used the same general approach instantiated by Navarro et al. (2016) and Recio et al. (2015) — namely comparing the effects of exposure schedule between a group of participants explicitly instructed to search for differences in the exposure phase, and a group receiving instructions which did not encourage self-supervision. The stimuli for discrimination were morphed look-alike faces that have been used in a number of previous studies of perceptual learning in my laboratory (e.g., Mundy et al., 2014; Mundy et al., 2009; Mundy et al., 2007). All participants (45 female, 3 male, Cardiff University undergraduates between 18 and 27 years of age) were tested for their ability to discriminate four face-pairs (see Table 1): prior to test one pair had received intermixed exposure, a second blocked exposure, a third had received exposure to the midpoint on the morph between the to-be-discriminated faces, while a fourth face pair was novel at the test stage (for details of the stimuli and the exposure phase, see: Mundy et al., 2007). The test phase consisted of same/different trials in which two stimuli from a face pair were presented in succession: on same trials, they were identical (e.g. A, A) and on different trials both stimuli from a given pair were presented (e.g. A, A*). The participants were asked to respond whether the stimuli were the same or different. The stimuli were presented for 0.5s with a 0.3s interval between them (filled with a pattern mask) and there were 16 trials (eight same, eight different) for each pair of stimuli (for further details of the test phase procedures see: Mundy et al., 2014; Mundy et al., 2009).

The key manipulation concerned the instructions given prior to the exposure phase. In Group Differences (n = 24), designed to promote self-supervision, participants were instructed:

Table 1. Design of Experiment

Condition	*Exposure*	*Same / Different Discrimination*
Intermixed	A, A*, A, A*, A, A*, A, A*, A, A*, A, A*	A vs A*
Blocked	B, B, B, B, B, B, B*, B*, B*, B*, B*, B*	B vs B*
Midpoint	C^m, C^m, C^m, C^m, C^m	C vs C*
Control	No exposure to D or D*	D vs D*

Note. A/A* to D/D* represent different face-pairs and C^m refers to the midpoint on the morph between faces C and C*. Both the assignment of face-pairs to condition and the order in which conditions were presented was counterbalanced.

> You are about to see a series of pairs of "lookalike" faces, please pay attention as the differences between them are subtle. You will be tested on your ability to discriminate between these "lookalike" faces in a later test phase.

In Group Attractiveness (n = 24), designed to minimise self-supervision, participants were instructed:

> You are about to see a series of faces. Please consider how attractive these faces are as you will be later asked to give an attractiveness score for these faces.

Apart from the instructions, both groups were treated identically. With respect to the statistical analysis, standard null-hypothesis significance testing does not directly assess whether the absence of a significant effect provides good evidence for there being no true relationship conditions. In contrast, Bayesian tests are based on calculating the relative probability of the null and alternative hypotheses, and thus afford the assessment of whether the evidence is in favour of either of these hypotheses. Therefore the analysis was performed both with classical ANOVA and Bayesian equivalents using JASP 0.7.1.12 (Love et al., 2009) with Bayes factors were calculated for factorial ANOVA as described by Rouder, Morey, Speckman and Province (2012) and Rouder, Morey, Verhagen, Swagman, and Wagenmakers (in press).[3]

3. The Bayes factor relates to the ratio of probability for the observed data under a model based on the null hypothesis compared to a model based on some specified alternative model. Bayes factors can be denoted as B_{01} when the data supports the null, or B_{10} when the data supports the alternative. The resulting Bayes factors can then be interpreted according to the

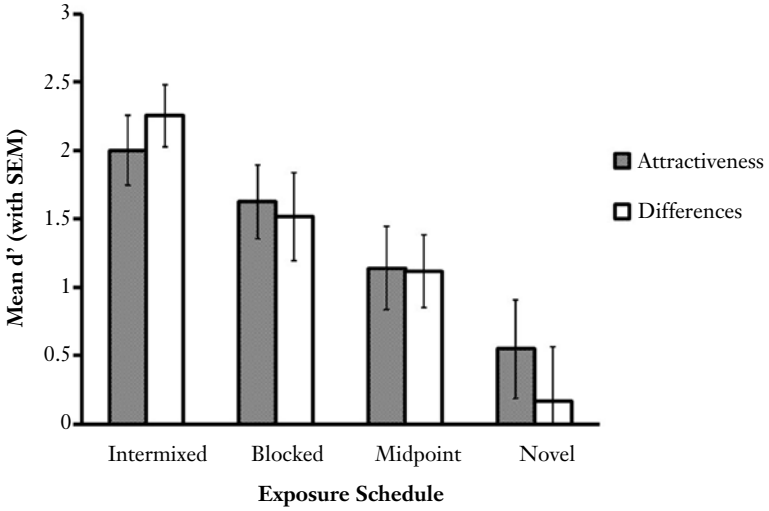

FIGURE 1. Shows mean d' scores (with SEM) for the discrimination test phase of the experiment as a function of instructions during pre-exposure ("consider attractiveness" vs "look for differences") and exposure schedule.

Figure 1 shows the data from the test phase. This clearly displays the usual exposure schedule effects, namely intermixed exposure produced superior discrimination to blocked or midpoint exposure, which in turn produced better performance than with novel stimuli. Critically, the effects of exposure schedule were equivalent in the group instructed to look for differences between stimuli and in the group instructed only to think about face attractiveness. This description of the results was confirmed by the ANOVA analysis which revealed a main effect of exposure schedule ($F(3,138) = 11.01, p < .001, B_{10} > 1000$), but no main effect of instruction group ($F(1,46) = 0.22, p = .664, B_{01} = 5.2$) or group by instruction interaction ($F(3,138) = 0.53$, p = .664, $B_{01} = 8.8$). Moreover, an analysis focused on only the intermixed and blocked exposure conditions also revealed an effect of exposure schedule ($F(1,46) = 4.10, p = .049, B_{10} = 2.4$), but no main effect of instruction group ($F(1,46) = 0.01, p = .937, B_{01} = 4.2$) or group by instruction interaction ($F(1,46) = 0.87, p = .356, B_{01} = 4.7$). In short, both

convention suggested by Jeffreys (1961) and recommended by Rouder, Speckman, Sun, Morey, and Iverson (2009): a Bayes factor between 1 and 3 gives anecdotal support, a factor between 3 and 10 suggests supporting evidence, while a factor of 10 indicates strong evidence.

classical and Bayesian analysis support the absence of any effect of instructions designed to promote or minimise implicit reinforcement through self-supervision even in the presence of the usual exposure schedule effects (both overall and focusing on the intermixed vs. blocked comparison).

Clearly, these new results stand in contrast to those reported by Navarro et al. (2016) and Recio et al. (2015). While it is unwise to draw firm general conclusions from such a small sample, several possibilities for this pattern of results do suggest themselves. Perhaps the most simple explanation is that the current "attractiveness" instructions were simply not as effective in preventing self-supervision as were the instructions used by Navarro et al. (2016) and Recio et al. (2015). It is not possible to rule out this explanation given the current data, although it is not immediately obvious why considering the general attractiveness of a group of faces would be less effective at preventing self-supervision than the instruction to simply look at the stimuli. A potentially more interesting possibility is that the difference in the pattern of results reflects the nature of the stimuli. If so, this would suggest that self-supervision and effortful searching for differences might be particularly prevalent when there are a small number of entirely diagnostic discriminative cues (such as the single unique feature added to the checkerboards) but that such processes are either ineffective or less prevalent when the there is no single defining feature that separates the stimuli (for an extended discussion of the issue of stimulus type in the context of effortful search effects, see: Jones & Dwyer, 2013).

While it may not be possible to come to a conclusion about the generality of self-supervision effects from the current small sample of studies which have used instruction to directly manipulate the tendency for human participants to self-supervise, these studies do demonstrate the potential for such an approach, and it can only be hoped that they will inspire more research along these lines.[4] But in the absence of further work of this type, it is worth considering whether there are other, more indirect means to assess the prevalence of effortful self-supervision effect in perceptual learning. One such approach is to consider the implications of studies seeking to examine the brain mechanisms underpinning perceptual learning.

4. In light of evidence that both perceptual learning (e.g. Leclercq, Cohen Hoffing, & Seitz, 2014; McDevitt, Rokem, Silver, & Mednick, 2014) and face processing (Godard & Fiori, 2010; Heisz, Pottruff, & Shore, 2013) may display male/female differences, such future research may want to include an explicit comparison of male and female subjects.

Considering the Brain

The psychophysical tradition of research has produced a wealth of evidence concerning the brain mechanisms involved in perceptual learning. There is abundant behavioural evidence that the enhanced discriminability produced by experience with visual stimuli is typically restricted to the stimulus orientation and retinal position used in training, and does not transfer to situations in which these were changed (e.g., Ball & Sekuler, 1982; Fiorentini & Berardi, 1980; Poggio, Fahle, & Edelman, 1992). Given that neurons with the requisite spatial resolution, location and orientation specificity are found only in primary visual cortex, this implies a critical role for primary sensory cortex in perceptual learning. That said, under appropriate training methods, perceptual learning can transfer across changes in location, stimulus orientation, and task (e.g., McGovern, Webb, & Peirce, 2012; Xiao et al., 2008; Zhang et al., 2010) implying that more central brain mechanisms are also involved. Moreover, functional neuroimaging studies of perceptual learning with a variety of stimuli and tasks (e.g., Lewis, Baldassarre, Committeri, Romani, & Corbetta, 2009; Mukai et al., 2007) have implicated the simultaneous involvement of primary visual cortex and higher brain regions including the frontal and supplementary eye-fields and dorsolateral pre-frontal cortex (regions that have been identified as part of a dorso-frontal attentional network: Corbetta & Shulman, 2002). Most critically for the current concerns, perceptual learning based on task irrelevant cues (which are not subject to self-supervision effects) has been identified with primary visual cortex mechanisms, while perceptual learning with task relevant cues has been linked to the actions of higher brain regions including the dorso-frontal attentional network noted above (Shibata, Sasaki, Kawato, & Watanabe, 2013; Watanabe & Sasaki, 2015).

Although none of the studies noted in the previous paragraph examined the sort of exposure schedule effects that are the central topic of this paper, the idea that task-irrelevant perceptual learning has been linked to primary cortex while task-relevant perceptual learning has been linked to higher cortical regions, suggests a possible way to interrogate the studies that have examined exposure schedule. Namely that if exposure schedule perceptual learning effects involve primary visual cortex then this would imply that they might be based on mechanisms independent of self-supervision, but if exposure schedule effects involve higher cortical and attentional regions

then it would imply effortful mechanisms that may well rely on self-supervision.

There have been two studies of which used fMRI methods to examine the functional brain mechanisms engaged by intermixed and blocked exposure schedules. The first (Mundy et al., 2009) used morphed faces and random checkerboards as stimuli, while the second (Mundy et al., 2014) used morphed faces, virtual reality scenes and random-dot patterns. Moreover, both studies used instructions which would have encouraged participants to inspect the stimuli in a way which would support self-supervision. The results of both studies were entirely in accordance with each other: irrespective of the type of stimuli involved, contrasting the brain regions activated after intermixed and blocked exposure revealed the involvement of *both* visual cortex and higher attentional regions (for an extended discussion of these results see, Dwyer & Mundy, 2016). Thus, with respect to the question posed above, the involvement of higher attentional regions implies that self-supervision might well contribute to the intermixed vs. blocked schedule effect with these stimuli, while the involvement of visual cortex implies that genuinely unsupervised mechanisms might also contribute to the effects of exposure schedule. Of course, Mackintosh would have warned that the mere correlation of different patterns of brain activity with different forms of perceptual learning does not directly demonstrate either the causal contribution of this activity to perceptual learning nor directly establish the psychological mechanisms involved. Notwithstanding such caveats, the involvement of both basic visual cortex and higher attentional brain structures in the difference between intermixed and blocked exposure schedules is at least suggestive with respect to the mechanisms involved.

Conclusions

In 2009 Mackintosh argued that the absence of explicit feedback in many perceptual learning experiments with humans did not constitute evidence of genuinely unsupervised learning mechanisms because of the possibility for self-supervision. In raising this (then hypothetical) possibility, Mackintosh questioned whether there was indeed any need to posit unsupervised learning mechanisms to account for any of the results — most critically those relating to the schedule of exposure to stimuli — which might be affected by

self-supervision. Mackintosh's caution was, as ever, perspicacious because subsequent studies have provided direct evidence for effortful search and self-supervision during the exposure phase of some studies which lack explicit feedback. However, other studies have shown evidence for perceptual learning with sub-threshold or task-irrelevant cue presentation, indicating that at least some examples of perceptual learning can occur without the contribution of self-supervision mechanisms. Moreover, both the new data reported here and a consideration of previous fMRI-based studies point towards the idea that genuinely unsupervised mechanisms also contribute to the effects of exposure schedule under some circumstances. So, Mackintosh's challenge has been met — but only in part. Unsupervised learning does appear to be important in human perceptual learning, and a number of mechanisms arising from the study of animal experiments have been proposed to explain it (for reviews of these potential mechanisms see, Mackintosh, 2009; Chris, Mitchell, & Hall, 2014). But the potential and actual contribution of self-supervision means that the currently available studies do not determine whether the possible mechanisms identified through the study of associative learning reflect the true underpinnings of unsupervised perceptual learning in humans. As ever, the last word should be for Nick: "Perceptual learning, like virtually every other interesting example of a psychological phenomenon, is surely multiply determined" (Mackintosh, 2009, p. 124).

References

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual-motion discrimination. *Science*, *218*(4573), 697-698.

Bennett, C. H., & Mackintosh, N. J. (1999). Comparison and contrast as a mechanism of perceptual learning? *Quarterly Journal of Experimental Psychology*, *52B*, 253-272.

Blair, C. A. J., & Hall, G. (2003). Perceptual learning in flavor aversion: Evidence for learned changes in stimulus effectiveness. *Journal of Experimental Psychology: Animal Behavior Processes*, *29*(1), 39-48.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201-215.

Dwyer, D. M. (1999). Retrospective revaluation or mediated conditioning? The effect of different reinforcers. *Quarterly Journal of Experimental Psychology*, *52B*, 289-306.

Dwyer, D. M. (2000). Formation of a novel preference and aversion by simultaneous activation of the representations of absent cues. *Behavioural Processes*, *48*, 159-164.

Dwyer, D. M. (2001). Mediated conditioning and retrospective revaluation with LiCl then flavour pairings. *Quarterly Journal of Experimental Psychology*, *54B*, 145-165.

Dwyer, D. M. (2003). Learning about cues in their absence: Evidence from flavour preferences and aversions. *Quarterly Journal of Experimental Psychology*, *56B*, 56-67.

Dwyer, D. M., Bennett, C. H., & Mackintosh, N. J. (2001). Evidence for inhibitory associations between the unique elements of two compound flavours. *Quarterly Journal of Experimental Psychology*, *54B*, 97-107.

Dwyer, D. M., Hodder, K. I., & Honey, R. C. (2004). Perceptual learning in humans: Roles of preexposure schedule, feedback, and discrimination assay. *Quarterly Journal of Experimental Psychology*, *57B*, 245-259.

Dwyer, D. M., & Mackintosh, N. J. (2002a). Alternating exposure to two compound flavors creates inhibitory associations between their unique features. *Animal Learning & Behavior*, *30*(3), 201-207.

Dwyer, D. M., & Mackintosh, N. J. (2002b). Perceptual learning: Alternating exposure to two compound flavours creates inhibitory associations between their unique features. *Animal Learning & Behavior*, *30*, 201-207.

Dwyer, D. M., Mackintosh, N. J., & Boakes, R. A. (1998). Simultaneous activation of the representation of absent cues results in the formation of an excitatory association between them. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 163-171.

Dwyer, D. M., & Mundy, M. E. (2016). Perceptual learning: Representations and their development. In R. C. Honey & R. Murphy (Eds.), *The Wiley-Blackwell handbook on the cognitive neuroscience of associative learning*, *9*, 201-222.

Dwyer, D. M., Mundy, M. E., & Honey, R. C. (2011). The role of stimulus comparison in human perceptual learning: Effects of distractor placement. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(3), 300-307.

Dwyer, D. M., Mundy, M. E., Vladeanu, M., & Honey, R. C. (2009). Perceptual learning and acquired face familiarity: Evidence from inversion, use of internal features, and generalization between viewpoints. *Visual Cognition*, *17*(3), 334-355.

Espinet, A., Iraola, J. A., Bennett, C. H., & Mackintosh, N. J. (1995). Inhibitory associations between neutral stimuli in flavor-aversion conditioning. *Animal Learning & Behavior*, *23*, 361-368.

Fiorentini, A., & Berardi, N. (1980). Perceptual-learning specific for orientation and spatial-frequency. *Nature*, *287*(5777), 43-44.

Gaffan, D. (1996). Associative and perceptual learning and the concept of memory systems. *Cognitive Brain Research*, *5*(1-2), 69-80.

Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology, 14*, 29-56.
Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appelton-Century-Crofts.
Gibson, E. J., & Walk, R. D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology, 49*, 239-242.
Gibson, E. J., Walk, R. D., Pick, H. L., & Tighe, T. J. (1958). The effect of prolonged exposure to visual patterns on learning to discriminate similar and different patterns. *Journal of Comparative and Physiological Psychology, 51*, 584-587.
Godard, O., & Fiori, N. (2010). Sex differences in face processing: Are women less lateralized and faster than men? *Brain and Cognition, 73*(3), 167-175.
Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*(2), 178-200.
Hall, G. (1991). *Perceptual and associative learning*. Oxford, England: Clarendon Press / Oxford University Press.
Hall, G., Blair, C. A. J., & Artigas, A. A. (2006). Associative activation of stimulus representations restores lost salience: Implications for perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes, 32*(2), 145-155.
Heisz, J. J., Pottruff, M. M., & Shore, D. I. (2013). Females scan more than males: A potential mechanism for sex differences in recognition memory. *Psychological Science, 24*(7), 1157-1163.
Honey, R. C., Bateson, P., & Horn, G. (1994). The role of stimulus comparison in perceptual learning: An investigation with the domestic chick. *Quarterly Journal of Experimental Psychology, 47B*, 83-103.
Jeffreys, H. (1961). *Theory of probability (3rd ed.)*. Oxford, UK: Oxford University Press / Clarendon Press.
Jones, S. P., & Dwyer, D. M. (2013). Perceptual learning with complex visual stimuli is based on location, rather than content, of discriminating features. *Journal of Experimental Psychology: Animal Behavior Processes, 39*(2), 152-165.
Jones, S. P., Dwyer, D. M., & Lewis, M. B. (2015). Learning faces: Similar comparator faces do not improve performance. *PLoS ONE, 10*(1), doi:10.1371/journal.pone.0116707.
Lavis, Y., & Mitchell, C. (2006). Effects of preexposure on stimulus discrimination: An investigation of the mechanisms responsible for human perceptual learning. *Quarterly Journal of Experimental Psychology, 59*(12), 2083-2101.
Leclercq, V., Cohen Hoffing, R., & Seitz, A. R. (2014). Uncertainty in fast task-irrelevant perceptual learning boosts learning of images in women but not men. *Journal of Vision, 14*(12). doi:10.1167/14.12.26.
Lewis, C. M., Baldassarre, A., Committeri, G., Romani, G. L., & Corbetta, M. (2009). Learning sculpts the spontaneous activity of the resting human brain. *Proceedings*

of the National Academy of Sciences of the United States of America, *106*(41), 17558-17563.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Wagenmakers, E.-J. (2009). JASP (Version 0.7). https://jasp-stats.org/.

Mackintosh, N. J. (1974). *The psychology of animal learning*. Oxford, England: Academic Press.

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

Mackintosh, N. J. (2009). Varieties of perceptual learning. *Learning & Behavior*, *37*(2), 119-125.

Mackintosh, N. J., Kaye, H., & Bennett, C. H. (1991). Perceptual learning in flavour aversion conditioning. *Quarterly Journal of Experimental Psychology*, *43B*, 297-322.

McDevitt, E. A., Rokem, A., Silver, M. A., & Mednick, S. C. (2014). Sex differences in sleep-dependent perceptual learning. *Vision Research*, *99*, 172-179.

McGovern, D. P., Webb, B. S., & Peirce, J. W. (2012). Transfer of perceptual learning between different visual tasks. *Journal of Vision*, *12*(11). doi:410.1167/12.11.4.

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 102-130). Oxford, England: Clarendon Press / Oxford University Press.

Mitchell, C., & Hall, G. (2014). Can theories of animal discrimination explain perceptual learning in humans? *Psychological Bulletin*, *140*(1), 283-307.

Mitchell, C., Kadib, R., Nash, S., Lavis, Y., & Hall, G. (2008). Analysis of the role of associative inhibition in perceptual learning by means of the same-different task. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*(4), 475-485.

Mitchell, C., Nash, S., & Hall, G. (2008). The intermixed-blocked effect in human perceptual learning is not the consequence of trial spacing. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*, 237-242.

Mukai, I., Kim, D., Fukunaga, M., Japee, S., Marrett, S., & Ungerleider, L. G. (2007). Activations in visual and attention-related areas predict and correlate with the degree of perceptual learning. *Journal of Neuroscience*, *27*, 11401-11411.

Mundy, M. E., Downing, P. E., Honey, R. C., Singh, K. D., Graham, K. S., & Dwyer, D. M. (2014). Brain correlates of experience-dependent changes in stimulus discrimination based on the amount and schedule of exposure. *PLoS ONE*, *9*(6), doi:10.1371/journal.pone.0101011.

Mundy, M. E., Dwyer, D. M., & Honey, R. C. (2006). Inhibitory associations contribute to perceptual learning in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(2), 178-184.

Mundy, M. E., Honey, R. C., Downing, P. E., Wise, R. G., Graham, K. S., & Dwyer, D. M. (2009). Material-independent and material-specific activation in functional MRI after perceptual learning. *Neuroreport*, *20*(16), 1397-1401.

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(2), 124-138.

Navarro, A., Arriola, N., & Alonso, G. (2016). Instruction-driven processing in human perceptual learning. *Quarterly Journal of Experimental Psychology*, *69*(8), 1583-1605.

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual-learning in visual hyperacuity. *Science*, *256*(5059), 1018-1021.

Recio, S. A., Iliescu, A. F., Mingorance, S. P., Bergés, G. D., & Hall, G. (2015). *The effect of instructions on perceptual learning using complex visual stimuli in humans*. Paper presented at the XXVII International Congress of the Spanish Society for Comparative Psychology (SEPC), Seville.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356-374.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (in press). Bayesian analysis of factorial designs. *Psychological Methods*.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225-237.

Shibata, K., Sasaki, Y., Kawato, M., & Watanabe, T. (2013). Perceptual learning is associated with different types of plasticity at different stages. *Journal of Vision*, *13*, 604. doi:10.1167/13.9.604.

Symonds, M., & Hall, G. (1995). Perceptual learning in flavour aversion conditioning: Roles of stimulus comparison and latent inhibition of common elements. *Learning and Motivation*, *26*, 203-219.

Trobalon, J. B., Chamizo, V. D., & Mackintosh, N. J. (1992). Role of context in perceptual learning in maze discriminations. *Quarterly Journal of Experimental Psychology*, *44B*, 57-73.

Tsushima, Y., & Watanabe, T. (2009). Roles of attention in perceptual learning from perspectives of psychophysics and animal learning. *Learning & Behavior*, *37*(2), 126-132.

Wang, T., Lavis, Y., Hall, G., & Mitchell, C. J. (2012). Location and salience of unique features in human perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(4), 407-418.

Wang, T., & Mitchell, C. J. (2011). Attention and relative novelty in human perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(4), 436-445.

Watanabe, T., & Sasaki, Y. (2015). Perceptual learning: Toward a comprehensive theory. *Annual Review of Psychology*, *66*, 197-221.

Xiao, L. Q., Zhang, J. Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology*, *18*(24), 1922-1926.

Zhang, J. Y., Zhang, G. L., Xiao, L. Q., Klein, S. A., Levi, D. M., & Yu, C. (2010). Rule-based learning explains visual perceptual learning and its specificity and transfer. *Journal of Neuroscience*, *30*(37), 12323-12328.

An Application of a Theory of Attention (Mackintosh, 1975) to Psychopathy: Variability in the Associability of Stimuli

G. M. Aisbitt, R. A. Murphy
Department of Experimental Psychology,
University of Oxford, UK

Abstract. Mackintosh (1975) described variations in the associability of stimuli as an account of attention. Theories of psychopathy have suggested that attention plays a central role in the symptomology of this psychopathology. We argue that individual learning differences associated with psychopathological characteristics (e.g., callous-unemotional traits, fearless dominance and impulsivity) are accompanied by a disruption in associability, perhaps involving context processing and stimulus integration. We present experimental evidence, which conflicts with predictions of a current theory of attentional deficits in psychopathy (Impaired Integration Theory; Hamilton, Hiatt Racer, & Newman, 2015). Finally, we demonstrate how changes in associability are able to capture both attentional and emotional accounts of psychopathy, and provide a means by which to reconcile these accounts with theories of associative learning.

Goals of the Chapter

This chapter will begin by presenting evidence related to reversal learning, a well-documented feature of psychopathic impulsivity. These data will be interpreted using the recently proposed Impaired Integration Theory of psychopathy (Hamilton et al., 2015), with particular focus on the learning and the emotion and cognition principles. We will give a brief overview of Mackintosh's (1975) theory, its key principles and examples of supporting evidence from the animal and human literature, before the altered reversal data is interpreted using this alternative model. Predictions from each of these theories regarding shift learning will be discussed, and we will then present evidence from an experiment designed to test shift learning in relation to psychopath-

ic traits. We argue that the Impaired Integration framework could incorporate Mackintosh's principles in order to explain the present data, and, more importantly, to mechanise the concepts outlined in the Impaired Integration framework. A novel Psychopathy Attention Theory framework designed to incorporate the principles of the Impaired Integration Theory into an associative account will be proposed.

Reversal Deficits in Psychopathy

Learning about the consequences of actions, and updating behaviour on the basis of this learning enables adaptive interaction with the environment (e.g., Byrom, Msetfi, & Murphy, 2015). At a basic level, much of behaviour involves people learning to make responses and withholding them in order to achieve reward and avoid punishment (Chatlosh, Neunaber, & Wasserman, 1985). But more than the simple ability to acquire associations, people are often in situations in which they must flexibly update these associations by being sensitive to changes to the contingencies of responding (Newman, Patterson, & Kosson, 1987). However, there are considerable individual differences in this ability, and the apparent disruption of this ability is a defining feature of psychopathy, a complex disorder more broadly characterised by callous-unemotional traits and impulsive antisocial behaviours (Cleckley, 1982; Hare, 1999; Lilienfeld & Widows, 2005; Patrick, Fowles, & Krueger, 2009). In psychopathy, though the initial acquisition of associations seems relatively preserved, the ability to adjust or reverse these associations is altered (see Brazil et al., 2013; Budhani, Richell, & Blair, 2006; Mitchell, Colledge, Leonard, & Blair, 2002).

For example, Brazil and colleagues (2013) compared the performance of offenders with and without psychopathy on a simple go/nogo task. The task required learning to respond to two predictive cues and then updating learning as contingencies reversed. During the learning phase participants were required to respond to one of two shapes (blue and green triangles) for reward of a variable probability. Participants in the so-called *explicit* condition were provided instructions directing attention to the possibility that the contingencies might vary, whilst those in the so-called *implicit* condition were not. The results showed that the groups high on the psychopathy dimension acquired the initial associations, but showed differences in reversal, though only in the explicit condition. This pattern was interpreted as showing that

explicit instruction guided attention away from important context or background cue information necessary for performance.

These findings are somewhat typical of evidence relating to learning differences and psychopathy. One specific interpretation is that this evidence points to a basic insensitivity to stimulus-outcome associations that guide behaviour, and this is a central prediction of Blair's (2005) Integrated Emotion Systems model (IES). According to the IES and other affective accounts (e.g., the Fear Dysfunction Hypothesis; Lykken, 1995), psychopathic individuals have a selective processing deficit regarding outcomes, particularly if they have affective content. This impairment limits their goal-directed abilities (Moul, Killcross, & Dadds, 2012), and supposedly results in poor fear conditioning and poor passive avoidance learning (the ability to learn to withhold responding to avoid punishment; Lykken, 1995). With the explicit attentional focus in Brazil et al.'s task, participants were unable to learn the shifting contingencies because the goal-directed system was employed. This hypothesis is a form of the inverse hypothesis in attention, which suggests that attentional resources directed to one set of stimuli comes at the expense of other stimuli (e.g., Thomas, 1970).

An alternative account is that psychopathic individuals have overly selective attention, and so are unable to alter their acquired or dominant response in a reversal phase (Response Modulation Hypothesis; Gorenstein & Newman, 1980). Attentional accounts, such as this (Gorenstein & Newman, 1980), suggest that psychopathy is characterised by a processing deficiency or blindness for cues outside the focus of attention which is itself a further example of the use of the inverse hypothesis. However rather than suggesting that this is an attentional effect researchers have suggested that the mechanism involves poor context retrieval and encoding. The impairment is believed to limit psychopathic individuals' ability to use contextual information (Hoppenbrouwers, Van der Stigchel, Slotboom, Dalmaijer, & Theeuwes, 2015), particularly when this contextual information conflicts with the pursuit of a current goal (MacCoon, Wallace, & Newman, 2004; Newman, 1998). This attentional problem may be exacerbated when overt re-allocation of attention is required, such as in the explicit condition, as compared to when these processes are governed more automatically, such as in the implicit condition.

Whilst both the affective and attentional accounts have their strengths, neither seems to provide an adequate explanation of the data, nor are the theories specified sufficiently to understand the mechanism underlying the effect. For instance it is not clear whether the deficit involves learning about the

cue-outcome relation, the responding associated with the cues, the valence of the cues or the outcomes, or the involvement of irrelevant background cues and outcomes. For example, attentional accounts do not explain the situational specificity of psychopathy deficits, and emotional-cognitive accounts do not explain why psychopathic individuals perform poorly on nonaffective tasks (Hamilton et al., 2015). As such, the recent Impaired Integration theory (Hamilton et al., 2015) was developed to reconcile and integrate these affective and attentional accounts, as well as neurobiological accounts, of psychopathy.

Hamilton et al.'s (2015) Impaired Integration Theory

Central to the Impaired Integration Theory is the failure of psychopaths to rapidly integrate complex, multisensory information, a failure that results in a perceptual bottleneck. The Impaired Integration Theory suggests that learning in psychopathy is impaired due to shallow processing, which prevents the integration of past and present information. Here the shallow attentional processes interfere with updating of contingencies, resulting in the characteristic impulsivity and perseveration (Brazil et al., 2013; Newman et al., 1987; Sadeh & Verona, 2008). Also related to such behaviours is the disruption of emotion and cognitive processes, resulting in a reduced propensity to process peripheral information. In turn, this is also believed to affect learning by diminishing the outcome. Therefore, the Impaired Integration Theory represents a model in which shallow perceptual processing, weak learning and poor emotional processing affect topographical representations and contribute to psychopaths' characteristic behaviours. At one level though this model is a simple re-characterisation of the effects described by the previous theories.

Importantly, the mechanism by which attention and learning interact is not specified. For instance, the impaired learning process is described as resulting from shallow information processing, which prevents the integration of new information with previous learning. However, which processes involved in information processing and learning are disrupted is not elaborated. Likewise the altered emotion and cognition process is said to reduce psychopaths' propensity, but not their ability, to process peripheral information. However, like the impaired learning process, the means through which these differenc-

es occur is not well outlined. As such, although these statements are useful in explaining the specificity of psychopathic abnormalities, such as impaired shift learning under some conditions, but not under others (Brazil et al., 2013), they do not provide insight in to the processes that cause these differences.

An alternative approach that does offer a mechanistic account involves interpreting the evidence of perseveration and reversal difference through Mackintosh's (1975) associative account of attention, the details of which will now be outlined. Currently there has been little attempt to harness theories of learning and attention as they have developed in the field of learning (although see Moul et al., 2012). One of the goals of this chapter is to show that the scientist in whose name this volume is dedicated, was establishing the principles that might be useful for the understanding of human variation in attention some 40 years ago.

Mackintosh's (1975) Theory of Attention

In his 1975 theory of attention, Mackintosh formalised a theory of selective attention by first establishing two principles of attention, gathered from an understanding of previous attention research. He then adapted a computational model of learning based on a standard Rescorla and Wagner (1972) learning algorithm, to capture these principles. The principles were 1) that we learn to attend to relevant and ignore irrelevant stimuli; and 2) that rather than assuming a type of reciprocal rule, in which attention is simply what happens if one is not attending to something else (the so called inverse rule), Mackintosh suggested a principle that the attention and associability of a cue are determined by the extent to which a cue is predictive of its consequences. He proposed that a stimulus-specific learning-rate parameter embedded within a standard error prediction algorithm could capture this notion.

While the first principle seems quite uncontroversial the second principle seems different in character from the bottleneck type models of attention proposed previously. Theories of selective attention had assumed that attention tunes in relevant stimuli at the expense of other stimuli, but Mackintosh allowed predictive or statistical information to determine attention.

Previous theories of associative learning had focused on the error term ($\lambda - V_A$), which represented the discrepancy between the expected outcome (V_A) and the actual outcome (λ; e.g., Rescorla & Wagner, 1972). Mackintosh's

theory incorporated this error prediction term, but also included α, a measure of how easily a stimulus is learnt about, referred to as its associability. Mackintosh proposed that associative change (ΔV) was governed by:

$$\Delta V_A = S\alpha_A (\lambda - V_A)$$

where V_A is the associative strength of stimulus A, *S* is a learning rate parameter, α_A is the associability of stimulus A and λ is the maximum conditioning possible to the outcome. The important addition was that α reflects the predictiveness of a cue correlation between the stimulus and an outcome, meaning that α would increase for a stimulus that was well correlated with its consequences, and α would decrease if a stimulus were irrelevant, or poorly correlated with its outcome. As such, Mackintosh used α to represent the idea that subjects learn to attend to relevant stimuli, and ignore irrelevant stimuli.

Importantly, this associability is relative, meaning that it depends in part on the associability of the other stimuli also present. This effect of other cues present is not, however, governed by the inverse hypothesis (Thomas, 1970), whereby a limited amount of associability is competed for by the stimuli present. Instead, the impact of other stimuli present on a stimulus' α is related to its relative validity, or its unique predictiveness. For example, a stimulus that is moderately correlated with the outcome may have high associability if it is the best predictor of the outcome, but may have low associability if another cue present is a better predictor (Hall, Mackintosh, Goodall, & Martello, 1977; Wagner, 1969). This is not because the presence of another cue 'uses up' some of the associability, but because the presence of another, more predictive cue, devalues the predictiveness of the original cue. This process of determining α was dependent on the following relations:

$$\Delta\alpha_A > 0 \text{ if } |\lambda - V_A| < |\lambda - V_X|$$

$$\Delta\alpha_A < 0 \text{ if } |\lambda - V_A| \geq |\lambda - V_X|$$

where V_x is the sum of the associative strength of all stimuli other than A present on that trial. As such α_A, increases if the outcome of a trial is predicted better by A than by all other stimuli present, and decrease if the outcome is better predicted by the other stimuli present.

Unlike the Rescorla-Wagner model, the use of a separable error term for each cue means that Mackintosh's (1975) model can only account for conditioned inhibition, in which excitatory cues (e.g., X+) facilitate the development of inhibitory strength to another cue with which it is not reinforced in compound (e.g., AX–; Le Pelley, 2004), with certain assumptions about the other parameters of the model. For example, inhibitory learning is predicted by assuming that lambda or beta are determined by the nature of the bidirectionality of the outcomes. In animal studies the presence and absence of the outcome is represented by values of lambda of 1 and 0 (for presence and absence). However in many human experiments the outcomes are more variable. For instance, in a category learning experiment in which the outcome is an increase or decrease in the outcome then lambda is logically defined symmetrically as +1 and –1 (Murphy et al., 2011), furthermore the effectiveness of these two outcomes might be different and represented by differences in beta. An alternative perspective was suggested by Le Pelley (2004), who incorporated a summed error term into his 'extended Mackintosh Model', given by:

$$\Delta V_A = S\alpha_A [\lambda - \Sigma V - \Sigma \bar{V})]$$

where ΣV is the summed associative strength of all cues present, and $\Sigma \bar{V}$ is the summed associative strength of all presented stimuli for the US representation.

Though Mackintosh's work drew from the human attention literature, his theory focused on explaining behaviours and phenomenon in the animal literature (Hall et al., 1977; Mackintosh & Holgate, 1968). However, the principle of associability varying with a subject's experience (e.g., correlation between stimulus and reinforcement) has also been demonstrated in humans. For example, Le Pelley et al. (2010) showed that human subjects can learn that stimuli are good predictors of their consequences. The idea that these stimuli might subsequently more easily enter into new associations because of higher associability was tested in a second Phase 2 of the experiment. Compared with poor predictors, good predictors were more easily acquired during a subsequent learning phase. That is to say, if a cue previously trained as predictive of outcome 1 (A1) was paired with a cue that was poorly predictive of outcome 1 (B), and the two were then trained in compound to predict a new outcome 2 (AB2), the previously predictive cue acquired greater associative strength than the

previously non-predictive cue. Demonstrations such as this provide evidence that Mackintosh's theory governs attentional processes in learning.

For the present purposes we suggest that it might be this associability related to attention that is disrupted in psychopathy. This hypothesis would suggest two novel ideas, the first is that there is a general associability problem which prevents a cue's associability from being updated, the second is that previous evidence for an affective component to the disorder may simply reflect the associability mechanism. Therefore, it is possible to account for reversal differences in psychopathic individuals by suggesting that their ability to determine a cue's associability is impaired.

Experiment

We sought to test whether psychopathic traits relate to subjects' ability to acquire and shift associations to excitatory and inhibitory cues. We used a probabilistic go/nogo task modified to incorporate participants' ratings of the cues on each trial. As shown in Table 1, in Phase 1, stimulus A was trained as an inhibitor (AX–, X+). In Phase 2, the strength of A's inhibitory association was weakened (A–, AY+). The two phases were analysed in two blocks (1 and 2) each consisting of four trials. This was done to capture the change in associative strength that occurred over the course of the phase.

Each of the theories that we have presented makes predictions about the ability of psychopathic and/or control individuals to acquire and shift these associations of A. These predictions are outlined below:

1) Predictions following Mackintosh's (1975) Theory

Mackintosh's theory predicts that α_A will be high in Phase 1 and so will facilitate acquisition of an inhibitory association. In Phase 2, Mackintosh's model predicts that there will be a decrease in α_A because it is inconsistently paired with the outcome. However, at the start of Phase 2, α_A will still be high given its predictive strength in Phase 1, and so will initially facilitate learning about the altered contingency between A and the outcome (Mackintosh & Holgate, 1968). As the trials progress α_A will decrease, and so learning will diminish. Therefore, Mackintosh's model predicts the low psychopathy participants should acquire inhibitory strength in Phase 1, and that this will shift in Phase 2, but that learning will decrease throughout Phase 2. If, as is suggest-

Table 1. Experimental Design

Phase 1	*Phase 2*
AX– X+	AY+ A–
BV+ V–	BW– B+
BV+	CZ– C–
AX–	DS+ D+

Note. Overall reinforcement direction is shown: excitatory (+) or inhibitory (–). A was the test cue. All other cues are included to counterbalance the training of A and inhibitory-excitatory exposure.

ed here, high psychopathy trait individuals' associability is disrupted, then α_A will not be reduced in Phase 2, and so shift learning will continue throughout Phase 2. Therefore a disruption of associability updating will be consistent with enhanced learning in the second phase.

2) Predictions following Hamilton et al.'s (2015) Impaired Integration Theory
For the high psychopathy participants, the Impaired Integration theory predicts that as A is central to the current goal-focus in Phase 1, A will acquire inhibitory strength. However, this acquisition of inhibitory strength will occur under conditions akin to those of high perceptual load in controls. As such, only shallow information processing will occur. This means that in Phase 2, high psychopathy participants will be unable to update their learning to A. This will result in the characteristic psychopathic perseveration, and participants will not show significant shift learning to A.

Participants were recruited from the community in a mixed sample of students and non-students. They were recruited on the basis of their not being colour-blind, given the nature of the task stimuli. Forty-seven individuals (24 male, 23 female; age M = 21.43 years, SD = 2.46) participated for course credit or payment.

The task required participants to learn when to respond and when to inhibit their responses to stimuli that had positive (excitatory) or negative (inhibitory) contingencies. Participants' learning was assessed through cue-ratings they made on every trial. The stimuli were consistent with the selected cover story of monitoring sales of a company following different products for sale. Each product for sale was represented by a word and colour. Each trial began with the presentation of a stimulus, which was either a product or a com-

pound of two products on the screen for 500ms. After 500ms an arrow appeared alongside the stimulus for a maximum of 1000ms, though the cues were removed and the trial progressed if participants responded (by pressing the "b" key).

The arrow was either: a green arrow pointing up accompanied by a high tone, signalling the need to respond; or a red arrow pointing down accompanied by a low tone, signalling the need to withhold the response. Red and green arrows were used so as to make the task more closely resemble standard go/nogo assessments in which the go and nogo stimuli are discriminable on the basis of a single feature, such a colour or letter type (Falkenstein, Hoormann, & Hohnsbein, 1999; Menon, Adleman, White, Glover, & Reiss, 2001).

The outcome (arrow) was probabilistic, occurring in one direction on 87.5% of trials for a given cue. Each stimulus was presented eight times in either the acquisition phase or the shift phase, meaning that a stimulus was reinforced in one direction on 7/8 trials, and in the other direction 1/8 trials. After a response was made, or the maximum time reached, participants were shown the stimulus and asked to rate it on a +1/–1 increment scale from –100 (inhibitory) to +100 (excitatory). Where a compound stimulus had been presented, one side of the image was covered over and participants rated the remaining image before this was reversed and they rated the other part of the compound.

Performance on this novel go/nogo task was compared to participants' levels of psychopathic traits, which were assessed using the Psychopathic Personality Inventory-Revised (PPI-R; Lilienfeld & Widows, 2005). This self-report questionnaire measured psychopathic traits using 154 items (sixty-four of which are reverse scored). Each item was rated on a scale of: false, mostly false, mostly true or true. The PPI-R is comprised of three factors: Self-Centred Impulsivity (SCI), consisting of the Machiavellian Egocentricity, Rebellious Non-Conformity, Blame Externalization, and Carefree Non-Planfulness subscales; Fearless Dominance (FD), comprising the Social Influence, Fearlessness and Stress Immunity subscales; and Coldheartedness (C). T-transformations were carried out on the PPI-R total scores and the score for each of the three factors (SCI, FD and C) on the basis of sample type (community vs. forensic), sex and age group of the participant. The internal consistency of the PPI-R was very good ($\alpha = .91$), and participants' means and standard deviations for the overall psychopathy score, and the SCI, FD and C scores are shown in Table 2.

Table 2. Participant Psychopathy *T*-Scores

	M	*SD*
Total PPI-R	50.74	10.24
Self-Centred Impulsivity	52.28	9.28
Fearless Dominance	48.85	9.59
Coldheartedness	50.49	12.39

Note: N = 47. The mean *t*-score is 50.

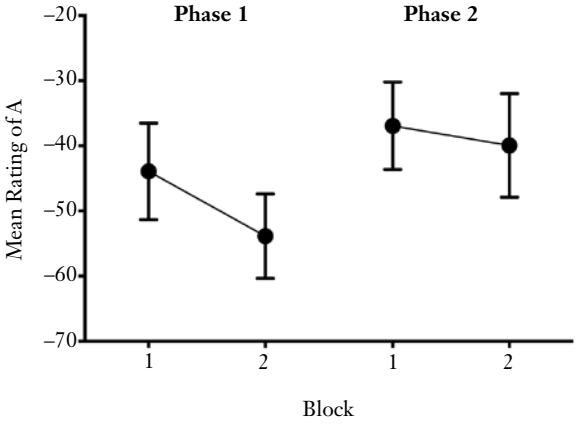

Figure 1. Participants' mean ratings of A in the first and second blocks of Phase 1 and 2. Overall participants showed significant shift learning across the phases. Participants continued to learn throughout Phase 1, but did not do so throughout Phase 2. Error bars represent 95% confidence intervals.

As shown in figure 1, at the end of each phase participants had shown significant shift learning to A [$t(46) = -3.79$, $p < .001$, 95% CI (–.18, –1.49)]. Consistent with Mackintosh (1975) participants showed significant learning throughout Phase1 [$t(46) = 2.83$, $p = .007$, 95% CI (.01, .33)], but not throughout Phase 2, suggesting initial shift learning that did not continue throughout the phase.

Also consistent with the predictions from Mackintosh's Theory, and inconsistent with the Impaired Integration Theory, psychopathic traits were relat-

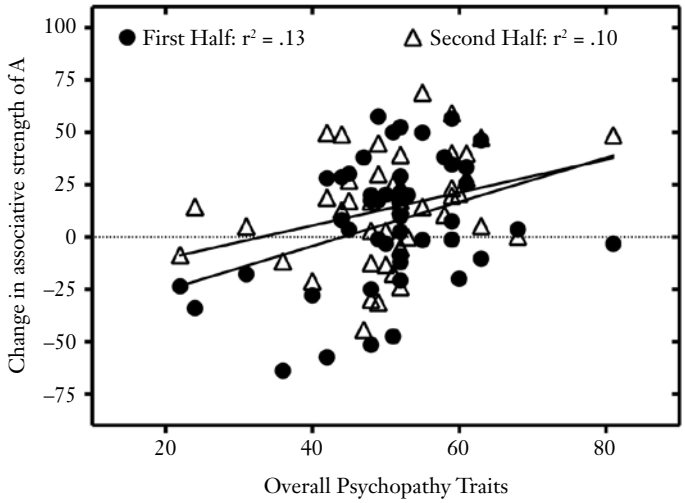

FIGURE 2. Participants' shift learning was significantly related to their psychopathic traits when comparing change in ratings across both the first and second half of the two phases.

ed to enhanced shift learning. As shown in figure 2, simultaneous multiple linear regression analysis showed that this shift learning was significantly predicted by overall psychopathic traits across both blocks 1 [$R^2_{adjusted} = .105$, $F(1,46) = 6.42$, $p = .015$, 95% CI (.25, .61)] and 2 [$R^2_{adjusted} = .083$, $F(1,46) = 5.15$, $p = .028$, 95% CI (.15, .53)].

Overall participants' performance was in line with the predictions of Mackintosh (1975), and they showed strong learning to A across Phase 1, but only showed initial shift learning to A when α_A was still high, but did not show learning across Phase 2 as α_A decreased. This shift learning was positively related to psychopathic traits: subjects' with higher psychopathic traits showed greater shift learning than those with lower psychopathic traits. This finding is inconsistent with the prediction of the Impaired Integration Theory that perseveration would occur due to shallow information processing preventing the updating of information about A in Phase 2. Therefore, psychopathic traits appear to be consistent with an impairment in adjusting α to reflect a cue's associability.

Novel Proposal: The Psychopathy Attention Theory

We argue that the broad features of the Impaired Integration Theory are compatible with Mackintosh's model, and that by integrating the two, a rule on which psychopathic learning is governed can be developed. There is considerable evidence that shallow information processing occurs in psychopaths for both affective (Blair, 2005; Budhani et al., 2006; Lykken, 1995) and non-affective stimuli (Hiatt, Schmitt, & Newman, 2004; Wolf et al., 2012). This shallow information processing, likened to a state of high perceptual load (Hamilton et al., 2015) is suggested to result in poor contextual processing (Hoppenbrouwers, Van der Stigchel, Slotboom, Dalmaijer, & Theeuwes, 2015; Aisbitt, Msetfi & Murphy, submitted). This demonstrates psychopaths' failure to use information outside of the current goal-pursuit to guide learning, which results in *impaired integration* of information.

According to Mackintosh's theory, a degree of cue competition governs learning, and a cue's associability is determined by its relative validity as a predictor of the outcome. This process of determining a cue's relative validity is dependent on integrating information about the predictiveness of all the other cues associated with the outcome. Therefore, if psychopathy is characterised by a problem in integrating information, this process of determining a cue's relative validity will be impaired. Specifically, we suggest that psychopathic traits adversely affect individuals' ability to compare predictiveness of a cue ($\lambda - V_A$) to the predictiveness of all the other cues ($\lambda - V_A$), a necessary step in marshalling changes in α. This is suggested to occur given the reduced processing capacity associated with psychopathic traits (Hamilton et al., 2015). We suggest that learning the predictiveness of the target cue is spared, but that psychopathic individuals differ in their perception of relative predictiveness of other cues. This is consistent with the overly goal-focused, central processing tendencies seen in psychopathic individuals at the expense of more peripheral or contextual stimuli (Baskin-Sommers, Curtin, & Newman, 2011; Wolf et al., 2012). As such, in psychopathic individuals the value of α will not adjust to reflect the cue's relative validity with the same degree of sensitivity that is seen in non-psychopathic individuals. We do not predict that this ability is binary, and instead suggest that psychopathic traits are inversely related to an individuals' ability to integrate information, an assertion consistent with the

present data showing a linear relationship between psychopathy and impaired associability.

We argue that this impaired ability is distinct from psychopaths' preserved use of error prediction terms to guide learning. There is substantial evidence of psychopathic individuals' ability to learn the about relationships between stimuli, responses and outcomes (Brazil et al., 2013; Budhani et al., 2006; Kiehl, Smith, Hare, & Liddle, 2000; Verona, Sprague, & Sadeh, 2012). It is accepted that psychopathic individuals do not struggle with this basic association formation (Moul et al., 2012), a process dependent on the accurate use of error predictions terms. However, differences arise with respect to psychopathic individuals' ability to reverse or shift these associations on the basis of new information. Historically, psychopathy was associated with response perseveration and an increased number of passive avoidance errors (Blair et al., 2004; Newman & Kosson, 1986; Newman, Patterson, Howland, & Nichols, 1990). But, more recent evidence suggests that this response reversal deficit is not all encompassing, and instead appears to be reliant on task parameters, such as an imbalance between excitatory and inhibitory trials (Kiehl et al., 2000), or whether an implicit or explicit learning condition is used (Brazil et al., 2013). It can easily be seen how these task parameters may influence α, and may, therefore, not involve an altered error term. For instance, an imbalance in excitatory trials that require a response and inhibitory trials that require the withholding of the response will influence the salience of the cues, and so the amount of attention they are allocated. As such, it is possible that data previously seen as evidence of an impaired error term are actually indicating the effects of altered associability.

We propose that these differences can be modelled by considering a trait factor modification to the modified Mackintosh model (Le Pelley, 2004; Mackintosh, 1975). We suggest that factor P is correlated with participants' level of psychopathic traits ($1<P<2$). We propose the following equation can be used to estimate the adjustment of α in relation to psychopathic traits:

$$\Delta\alpha_A > 0 \text{ if } |\lambda - V_A| < P|\lambda - V_X|$$

$$\Delta\alpha_A < 0 \text{ if } |\lambda - V_A| \geq P\ |\lambda - V_X|$$

whereby as psychopathic traits increase, P increases, and so consequently reduces the perceived predictiveness of other stimuli when determining the unique predictiveness of a given stimulus. This underestimation of the predictiveness of other cues relative to a given stimulus results in stimuli being perceived as more predictive than they necessarily are, and correspondingly α values that are too high.

As well as capturing the attentional components of psychopathy, we suggest that the parameter α is able to explain, at least in part, the emotional-affective component of psychopathy. Previous research has shown that the valence of an outcome can influence attention that is directed to cues associated with it, with participants paying more attention to cues that predict a high-value outcome as compared to cues that predict a low-value outcome (Le Pelley, Mitchell, & Johnson, 2013). This increased associability of cues that are predictive of a high-value outcome will be reflected through the cues' α. Therefore, if the changeability of the α parameter is disrupted, a high-value cue (either reward or punishment) will not be able to increase the ease with which psychopathic individuals learn about the cue. This would result in cues associated with high-value outcomes and those associative with low-value outcomes being learnt about at similar rates.

Evidence of this apparent insensitivity to the value of an outcome is seen in psychopaths' significant impairment when distinguishing between cues associated with varying degrees of both punishment and reward (Blair, Morton, Leonard, & Blair, 2006). Blair and colleagues used a task in which participants were required to choose between two objects associated with different levels of reward or punishment (+/– 100, 200, 400, 800, 1600). The psychopathy group learnt the association between the cues and their respective outcomes at similar rates, whilst the control group showed greater learning to cues associated with a higher positive or negative value. However rather than suggesting that it is the processing of affective information that is impaired in psychopaths, it is more parsimonious to consider that the mechanism for changes in associability is altered. Specifically, we suggest that the error prediction term does not receive influence from changes in associability. This may provide an a more general explanation for learning and attention deficits then have hitherto been suggested.

Conclusions

In this chapter we have presented our Psychopathy Attention Theory (PhAT) to explain the unique pattern of behavioural and cognitive differences seen in psychopathy. The equation incorporates principles of: 1) the recent Impaired Integration Theory that has attempted to reconcile attentional and emotional-cognitive accounts; and 2) Mackintosh's (1975) model of associability and associative learning. PhAT represents the idea that adjustment of a cue's associability is disrupted in psychopathy. This *fixed associability* results from a failure to integrate contextual information, and means that affective information does not alter the rate of learning (though the affective information itself can still be learnt about). We presented evidence in support of this explanation, and have discussed how this novel account can be used to explain other data, such as psychopaths' apparent insensitivity to affective outcomes. Therefore, we argue that PhAT provides a novel means by which to model the unique pattern of psychopathic behaviour and cognitions.

References

Blair, K. S., Morton, J., Leonard, A., & Blair, R. J. R. (2006). Impaired decision-making on the basis of both reward and punishment information in individuals with psychopathy. *Personality and Individual Differences*, *41*(1), 155-165. doi:10.1016/j.paid.2005.11.031.

Blair, R. J. R. (2005). Applying a cognitive neuroscience perspective to the disorder of psychopathy. *Development and Psychopathology*, *17*, 865-891. doi:10.1017/S0954579405050418.

Blair, R. J. R., Mitchell, D. G. V., Leonard, A., Budhani, S., Peschardt, K. S., & Newman, C. (2004). Passive avoidance learning in individuals with psychopathy: Modulation by reward but not by punishment. *Personality and Individual Differences*, *37*(6), 1179-1192. doi:10.1016/j.paid.2003.12.001.

Brazil, I. A., Maes, J. H. R., Scheper, I., Bulten, B. H., Kessels, R. P. C., Verkes, R. J., & de Bruijn, E. R. A. (2013). Reversal deficits in individuals with psychopathy in explicit but not implicit learning conditions. *Journal of Psychiatry & Neuroscience*, *38*(4), 13-20. doi:10.1503/jpn.120152.

Budhani, S., Richell, R. A., & Blair, R. J. R. (2006). Impaired reversal but intact acquisition: Probabilistic response reversal deficits in adult individuals with psychopathy. *Journal of Abnormal Psychology*, *115*(3), 552-558. doi:10.1037/0021-843X.115.3.552.

Byrom, N. C., Msetfi, R. M., & Murphy, R. A. (2015). Two pathways to causal control: Use and availability of information in the environment in people with and without signs of depression. *Acta Psychologica*, *157*, 1-12. doi:10.1016/j.actpsy.2015.02.004.

Chatlosh, D. L., Neunaber, D. J., & Wasserman, E. A. (1985). Response-outcome contingency: Behavioral and judgmental effects of appetitive and aversive outcomes with college students. *Learning and Motivation*, *16*(1), 1-34. doi:10.1016/0023-9690(85)90002-5.

Cleckley, H. (1982). *The Mask of Sanity* (revised edition). Mosby Medical Library.

Gorenstein, E. E., & Newman, J. P. (1980). Disinhibitory psychopathology: A new perspective and a model for research. *Psychological Review*, *87*(3), 301-15. doi: doi:10.1037/0033-295x.87.3.301.

Hall, G., Mackintosh, N. J., Goodall, G., & Martello, M. D. (1977). Loss of control by a less valid or by a less salient stimulus compounded with a better predictor of reinforcement. *Learning and Motivation*, *8*(2), 145-158. doi:10.1016/0023-9690(77)90001-7.

Hamilton, R. K. B., Hiatt Racer, K., & Newman, J. P. (2015). Impaired integration in psychopathy: A unified theory of psychopathic dysfunction. *Psychological Review*, *122*(4), 770-791. doi:10.1037/a0039703.

Hare, R. D. (1999). *The Hare Psychopathy Checklist-Revised: PCL-R*. MHS, Multi-Health Systems.

Hiatt, K. D., Schmitt, W. A., & Newman, J. P. (2004). Stroop tasks reveal abnormal selective attention among psychopathic offenders. *Neuropsychology*, *18*(1), 50-59. doi:10.1037/0894-4105.18.1.50.

Hoppenbrouwers, S. S., Van der Stigchel, S., Slotboom, J., Dalmaijer, E. S., & Theeuwes, J. (2015). Disentangling attentional deficits in psychopathy using visual search: Failures in the use of contextual information. *Personality and Individual Differences*, *86*(November), 132-138. doi:10.1016/j.paid.2015.06.009.

Kiehl, K. A., Smith, A. M., Hare, R. D., & Liddle, P. F. (2000). An event-related potential investigation of response inhibition in schizophrenia and psychopathy. *Biological Psychiatry*, *48*(3), 210-221. doi:10.1016/s0006-3223(00)00834-9.

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology*, *56*(1), 68-79. doi:10.1080/02724990244000179.

Le Pelley, M. E., Mitchell, C. J., & Johnson, A. M. (2013). Outcome value influences attentional biases in human associative learning: Dissociable effects of training and instruction. *Journal of Experimental Psychology: Animal Behavior Processes*, *39*(1), 39-55. doi:10.1037/a0031230.

Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., & Murphy, R. A. (2010) Stereotype formation: Biased by association, *Journal of Experimental Psychology: General*, *139*, 138-161.

Lilienfeld, S. O., & Widows, M. R. (2005). *Psychopathic Personality Inventory - Revised: Professional Manual*. Lutz, FL: Psychological Assessment Resources.

Lykken, D. T. (1995). *The antisocial personalities*. Psychology Press.

MacCoon, D. G., Wallace, J. F., & Newman, J. P. (2004). Self-regulation: Context-appropriate balanced attention. In *Handbook of Self-Regulation Research* (pp. 422-444). New York: Guildford Press.

Mackintosh, N. J., & Holgate, V. (1968). Effects of inconsistent reinforcement on reversal and nonreversal shifts. *Journal of Experimental Psychology*, *76*(1), 154-159. doi:10.1037/h0025310.

Mitchell, D. G. V., Colledge, E., Leonard, A., & Blair, R. J. R. (2002). Risky decisions and response reversal: Is there evidence of orbitofrontal cortex dysfunction in psychopathic individuals? *Neuropsychologia*, *40*(12), 2013-2022. doi:10.1016/S0028-3932(02)00056-8.

Moul, C., Killcross, S., & Dadds, M. R. (2012). A model of differential amygdala activation in psychopathy. *Psychological Review*, *119*(4), 789-806. doi:10.1037/a0029342.

Murphy, R. A., Schmeer, S., Vallée-Tourangeau, F., Mondragon, E., & Hilton, D. (2011). Making the illusory correlation appear and then disappear: The effect of more learning. *Quarterly Journal of Experimental Psychology*, *64*, 24-40.

Newman, J. P. (1998). Psychopathic behaviour: An information processing perspective. In *Psychopathy: Theory, Research and Implications for Society* (pp. 81-104). Springer Netherlands.

Newman, J. P., & Kosson, D. S. (1986). Passive avoidance learning in psychopathic and nonpsychopathic offenders. *Journal of Abnormal Psychology*, *95*(3), 252-256. doi:10.1037/0021-843x.95.3.252.

Newman, J. P., Patterson, C. M., Howland, E. W., & Nichols, S. L. (1990). Passive avoidance in psychopaths: The effects of reward. *Personality and Individual Differences*, *11*(11), 1101-1114. doi:10.1016/0191-8869(90)90021-i.

Newman, J. P., Patterson, C. M., & Kosson, D. S. (1987). Response perseveration in psychopaths. *Journal of Abnormal Psychology*, *96*(2), 145-148. doi:10.1037/0021-843x.96.2.145.

Patrick, C. J., Fowles, D. C., & Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: Developmental origins of disinhibition, boldness, and meanness. *Development and Psychopathology*, *21*(3), 913-938. doi:10.1017/S0954579409000492.

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

Sadeh, N., & Verona, E. (2008). Psychopathic personality traits associated with abnormal selective attention and impaired cognitive control. *Neuropsychology*, *22*(5), 669-680. doi:10.1037/a0012692.

Thomas, D. R. (1970). Stimulus selection, attention and related matters. In J. H. Reynierse (Ed.), *Current Issues in Animal Learning* (pp. 311-356). Lincoln: University of Nebraska Press.

Verona, E., Sprague, J., & Sadeh, N. (2012). Inhibitory control and negative emotional processing in psychopathy and antisocial personality disorder. *Journal of Abnormal Psychology*, *121*(2), 498-510. doi:10.1037/a0025308.

Wagner, A. R. (1969). Stimulus validity and stimulus selection in associative learning. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental Issues in Associative Learning* (pp. 90-122). Halifax: Dalhousie University Press.

Wolf, R. C., Carpenter, R. W., Warren, C. M., Zeier, J. D., Baskin-Sommers, A. R., & Newman, J. P. (2012). Reduced susceptibility to the attentional blink in psychopathic offenders: Implications for the attention bottleneck hypothesis. *Neuropsychology*, *26*(1), 102-109. doi:10.1037/a0026000.

*Bottom-up Associative Mechanisms and Generalization Can Account for Apparent Contrast Effects Between Causes of Different Strengths**

Janie Lober, A. G. Baker
Department of Psychology, McGill University,
Montreal, QC, Canada

Irina Baetu
School of Psychology, University of Adelaide,
Adelaide, SA, Australia

Abstract. The presence of a strong predictor of an outcome often reduces judgments of a weaker one. One top-down explanation of this finding has been that the stronger cause provides all the information that is needed to predict the outcome — so the weaker cause is discounted. We have reported results that are inconsistent with this view because judgments of a moderate cause can be enhanced rather than reduced. In two experiments we replicate the fact that the presence of highly informative causes of opposite polarity sometimes enhance rather than reduce judgments of moderate causes. Furthermore, stronger causes of the same polarity can even push judgments of a moderate cause past zero so the causes are judged as opposite to their objective polarity. Moreover, we find that manipulating the number of common perceptual elements in the causes or the salience of the context moderates these competition effects. We present simulations with the Rescorla-Wagner model that are more consistent with these effects than the top-down statistical model.

Introduction

In his seminal (1975) book on animal learning N. J. Mackintosh emphasized the role of associative learning, selective attention and possible cognitive

* This work was supported by an Australian Research Council Discovery Early Career Researcher Award to I. B. and a Natural Sciences and Engineering Research Council of Canada Discovery Grant awarded to A. B.

representational processes (for example in avoidance learning). Later one of us continued his original work on learned irrelevance, extending it to a hypothesis that animals actually represent the relationship between events in the environment (Baker, 1976; Baker & Machintosh, 1977). The work here explores similar ideas in the representation of events in human causal reasoning about two competing causal events.

There are two popular views as to how a person might judge the relationship between cause and effect. In the first, the reasoner is seen as an active problem-solver who observes the causal data, extracts regularities, and then applies causal rules to these data (e.g., Mitchell, De Houwer, & Lovibond, 2009; Waldmann, 1996). The reasoning may or may not be rational, in the sense that it may not accurately reflect reality, but it is rule-based and retrospective. The basic causal rules include: temporal order (causes come before effects), temporal contiguity (causes and effects occur close together in time), spatial contiguity (causes and effects are usually proximal), and physical similarity or appropriateness of cause and effect (heat will boil water but not move a billiard ball). An alternative view is that, when faced with causal data, any regularity might cause associations to form in an automatic bottom-up manner (e.g., Dickinson, Shanks, & Evenden, 1984). When asked, the reasoner's report about the cause is influenced by the strength of these associations.

At first glance, it would seem straightforward to distinguish between the rich rule-based account and the apparently impoverished — although we would argue more parsimonious — associative-based account. However, a moment's reflection reveals that the basic rules of causal models bear a remarkable similarity to the fundamental laws of association that came from the British empiricists (e.g., Hume, 1740). And, indeed, we, and many others, have demonstrated that simple associative nets such as the Rescorla-Wagner model (Rescorla & Wagner, 1972) and its modifications are sensitive to temporal order, contiguity and even contingency (e.g., Baetu & Baker, 2009, in press; Wasserman, Elek, Chatlosh, & Baker, 1993). It has been argued that, because these associations are semantically neutral, when causal information is ambiguous people require mental models to tell them which cause is appropriate to pair with which effect (Waldmann, 1996). However, the brain is not wired in a neutral manner so some associative links are more likely to form than others. For example, animals are likely to form an illness-induced aversion to gustatory cues (like a novel flavor) but not to physical cues such as a sound or a light (Garcia & Koelling, 1966). Thus, the fundamental rules,

laws, or models of causality do not provide a very good assay of which view of causality is more appropriate.

We have been interested in a more complex rule-based characteristic of causal discovery. Participants are asked to make inferences about two causes that are correlated with an effect. Generally one cause (the alternative cause A) is more highly correlated with the effect than the second (target X) cause. People usually attribute strong causal power to the alternative but ascribe little power to the target. It is judged to be much weaker than it would be judged when the alternative is not correlated with the outcome (Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Darredeau, Baetu, Baker, & Murphy, 2009). This process is often called blocking: the strong alternative "blocks" associations to the weaker target. Blocking occurs when both causes are generative of (positively correlated with) the effect or are preventive of (negatively correlated with) the effect (e.g., Baetu & Baker, 2012; Darredeau et al., 2009). It appears that this finding is inconsistent with a causal model because people ignore the objective correlation between the target and the effect.

However, it should be obvious that because the alternative and the target are correlated with the effect they can be correlated with each other. This is true in most causal learning experiments so when the information in the correlation between the stronger alternative and the effect is taken into account, the target cause provides little or no information (Baker, Murphy, & Vallée-Tourangeau, 1996; Spellman, 1996). If a causal candidate provides little information about the presence of the effect, then it is reasonable to assume that it is not a good candidate for causal power. Hence attributing low causal power to the target is not a "failure" to learn or represent the objective correlation, rather it is an accurate representation of a rational causal model; the more informative cause gets causal precedence and the target is only granted power if it adds extra information.

Cheng and others provided a mathematical justification of this rule called the probabilistic contrast model (PCM; Cheng & Novick, 1990; 1992). If a potential target cause, X, is sometimes followed by an effect, the contingency (ΔP_X) between X and E is computed by: $\Delta P_X = P(E|X) - P(E|\text{no } X)$. $P(E|X)$ is the probability of the effect (E) in the presence of cause X, and $P(E|\text{no } X)$ is the probability of E in the absence of X. This simple rule is not sufficient if there is a second potential cause. When X occurs in the presence of an alternative cause (A) the effect of X might be confounded with that of A. Thus, the contingency between X and E should be computed while controlling

for the presence or absence A — by holding A constant. That is $\Delta P_{X|A}$ or $\Delta P_{X|(no\ A)}$ is computed only on trials when A is present or absent. For example the contingency between X and E in the presence of A is $\Delta P_{X|A} = P(E|X, A) - P(E|no\ X, A)$. If X influences E independently of the effect of A, then the conditional contingency $\Delta P_{X|A}$ should not be null. Critically, in the previous blocking treatments (e.g., Baker et al., 1993; Baker, Vallée-Tourangeau, & Murphy, 2000; Darredeau et al., 2009; Vallée-Tourangeau, Murphy, & Baker, 1998), the contingencies $\Delta P_{X|A}$ and $\Delta P_{X|(no\ A)}$ were null. Thus, the probabilistic contrast rule can account for reduced perceived effectiveness of a target cause when a stronger alternative is present.

This rule held when both causes were generative, or were preventive, or when one was preventive and the other generative. That is, $\Delta P_{X|A}$ or $\Delta P_{X|(no\ A)}$ was null regardless of causal polarity (e.g., Baker et al., 1993; Baker et al., 2000; Darredeau et al., 2009; Vallée-Tourangeau et al., 1998). Information is information; and its polarity matters little. For example, a strong negative cause reduces the relative informativeness of a weaker correlated positive cause because the absence of the strong negative cause (A) signals the occurrence of the effect more reliably than the presence of the weak positive cause (X). There is evidence that this prediction sometimes holds (Baker et al., 1993; Baker et al., 2000). Nevertheless, it is also true that a simple associative model (Rescorla & Wagner, 1972) generates associations that map directly onto these predictions (Baker et al., 2000; Darredeau et al., 2009; Vallée-Tourangeau et al.,1998). So even the "sophisticated" information processing in blocking does not distinguish between the two views.

To complicate matters, we have carried out a series of experiments that were originally designed to investigate limits on how two highly correlated alternatives (A and B), that individually did not predict more outcomes than the target (X), reduce judgments of the target (Darredeau et al., 2009). The answers to this question are not germane here, but one finding is. While two strong competing causes of the same polarity did reduce judgments of the weaker target cause, when the causes had opposite polarities the apparent strength of the modest, generative or preventive, cause was enhanced rather than blocked. This enhancement is inconsistent with the PCM because strong causes correlated with the target (X) reduce the information provided by X and should reduce the impression of causal power. We argued that this result was consistent with the perceptual principle of contrast, where a bright visual field will make a moderate one appear darker and with behavioural con-

trast where a larger reward will make a smaller reward appear less valuable. Contrast implies that a strong cause will push judgments of a weaker cause away from it. For example, a strong preventive cause might make a modest generative cause seem stronger.

A subtler characteristic of our data was also consistent with this notion. With causes of the same polarity, the modest polarity target was often judged not only as weaker but to be of the opposite polarity. From the statistical information processing (e.g., PCM) point of view this was surprising because, for example, a strong positive A renders X weaker but not negative (Baker et al, 1996; Spellman, 1996). However, the observed reduction past zero arises simply from the contrast notion in which the strong cause could easily push the weaker one past zero (Darredeau et al., 2009).

This leaves us with the dilemma that sometimes a cause blocks judgments of a target of the opposite polarity (Baker et al., 1993; 2000) and at other times strengthens it (Darredeau et al., 2009, see also Vallée-Tourangeau et al., 1998). Although this finding poses problems for both the statistical models approach and some associative models, simulations that manipulate the salience of the causal context predict the weakening of blocking with causes of opposite polarity and sometimes its reversal (i.e., enhancement). Indeed, the Rescorla and Wagner (1972) model predicts blocking of a moderate target cause when a strong negative alternative cause is also present and contextual cues are included in the simulation. In contrast, when the context is omitted from the simulation (its salience is zero), the model predicts enhancement (i.e., higher associative strength) of the moderate target when a strong negative alternative is present. The model thus predicts blocking or enhancement with causes of opposite polarity depending on whether the context is assumed to be salient or not (figure 1). Including the context is a common and uncontroversial tool for associative modelers. The context is that set of cues that is always present, signaling a situation in which the effect might occur, but does not accurately predict the timing of the effect. For example, in animal learning the context is the conditioning chamber and other static cues present during training. Clearly contextual cues are somewhat predictive of the effect, but the "true" cause (e.g., a conditioned stimulus) is usually a better predictor and is attributed causal efficacy. Nonetheless, the context does compete with the cause for associative strength and can modulate its judged effectiveness.

The Rescorla-Wagner model's predictions regarding the target cause X when the alternative cause A is of opposite polarity critically depend on the sa-

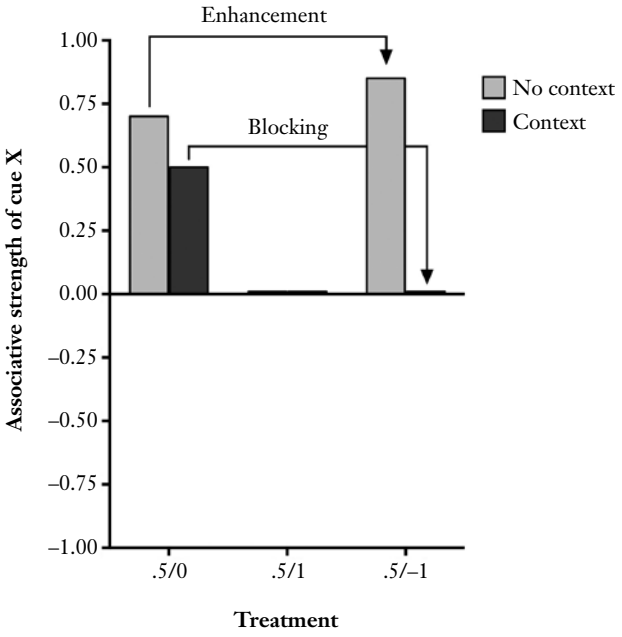

FIGURE 1. Simulations with the Rescorla-Wagner model (Rescorla & Wagner, 1972). The figure shows the asymptotic associative strength of a moderate target cause that has an overall contingency of 0.5 with the effect. In the first, control, treatment (.5/0), the moderate cause is paired with an alternative cause that is uncorrelated with the effect (i.e., the contingency between the alternative and the effect is null). In the second treatment (.5/1) the moderate cause is paired with a strong positive alternative that is perfectly correlated with the effect (the contingency between the alternative and the effect is 1). The model predicts 'blocking' of the target cause, as its predicted associative strength is reduced compared to the control (.5/0) treatment. In the third treatment (.5/–1) the moderate cause is paired with a strong negative alternative that is perfectly correlated with the absence of the effect (the contingency between the alternative and the effect is –1). The model predicts blocking of the moderate cause (compared to the control treatment) if the context is represented in the simulation (dark grey bars). However, if the context salience is set to zero (so it cannot accrue associative strength), the model predicts enhancement of the moderate cause (i.e., higher associative strength in treatment .5/–1 compared to treatment .5/0; light grey bars).

lience of the context. Figure 1 shows simulations for a treatment where X has a moderate positive overall contingency with the effect and A has a strong negative contingency (Treatment .5/–1, where the treatment label designates the overall contingencies of X and A, ΔP_X / ΔP_A). The trial context (i.e., those cues that signal a trial is occurring) plays an important role in Treatment .5/–1 because it is always paired with the effect when presented on its own (see Table 1). This is inevitable when the contingency of A is –1, as this effectively means that the effect occurs on all trials on which A is absent, including the context-alone trials. If the context is salient, it acquires more excitatory associative strength than X and eventually 'blocks' X because it acquires strength on context-alone trials, whereas X is unaffected. So the context blocks learning about X because it signals the outcome both when X is present (these are context-and-X+ trials; where + denotes the presence of the effect) and when X is absent (on context+ trials), rendering X a redundant predictor of the effect. Thus, with a salient context, the context, rather than A directly, prevents X from acquiring associative strength. If the context salience is zero, blocking in is not expected in Treatment .5/–1. Indeed with a zero context salience the model predicts an asymptotic enhancement effect, and with relatively low context salience, the model predicts a transient enhancement effect (see Darredeau et al., 2009).

Thus in these simulations, the reversal of the blocking effect critically depends on the salience of the causal context. We therefore tested whether manipulations of the context would influence the reversal, or at least reduce the magnitude, of the blocking effect. This context effect could potentially explain a number of contradictory findings with this treatment, as we have sometimes found blocking (e.g., Baker et al., 1993), and sometimes enhancement (e.g., Darredeau et al., 2009).

Breaking causes or any other stimulus into individual elements is also a common practice and dates at least as far back as Estes's stimulus sampling theory (Atkinson & Estes, 1963). It has its roots in the perceptual notion that any percept is not unitary but consists of many different but correlated inputs. This notion provides a simple account of generalization between different stimuli. The more common elements they possess, the more similar two stimuli will appear. Manipulating the ratio of common to unique elements of the two causes can also, in principle, influence the magnitude of the blocking effect or its reversal. This is because two causes that share many elements will be perceived as more similar, and experience with one of the two causes may generalize to the other, therefore reducing cue competition effects such as blocking.

In the present set of experiments we manipulated the strength of the causal context and the number of common elements possessed by competing causes. We did this for two main reasons. We wished to investigate how these simple associative manipulations would influence people's attributions of causal efficacy of a moderate target. As well, we wondered if the context manipulation could reduce and perhaps even reverse the enhancement effect with cross-polarity causes.

Experiment 1

The objective of the first experiment was to attempt to replicate the basic finding that, while strong causes of the same polarity usually block judgments of moderate causes, strong causes of the opposite polarity can enhance judgments of moderate causes. Moreover, the experiment was designed to see whether generalization engendered by the presence of common elements in the competing causes and/or the presence of a more salient context, generated by adding a constant element present on all trials, would reduce the enhancement and blocking generated by competing causes of the same or opposite polarity. We used a new scenario in which the participants pretended they were in a spaceship searching for alien life forms on different planets. They observed a display with five on-off indicators indicating environmental conditions at various sites on the planet. They observed the indicators and were then told if a life form was present at that location. The indicators represent causal cues, whereas the presence of a life form represents the effect.

The indicators represent different elements of the overall environmental state at the location. The similarity of the target (X) and alternative (A) cause was manipulated via a common element: when A and X shared a common element they were more similar to each other than when they did not share any common elements. The target cause (X) and the competing cause (A) could each be represented by one or two indicators. If there were no common elements, A and X each consisted of a single unique indicator light. If there was a common element, A and X each consisted of two indicator lights: one light was unique to each cause and the other was common to A and X. We also manipulated the salience of the trial or observation context. A salient context was represented by an indicator light, representing a constant environmental factor, that was present on all observations. For the less salient

Table 1. The frequency of events for each contingency treatment.

Type of trial and contingencies	*.5/0 (control)*	*.5/1 (same polarity)*	*.5/–1 (opposite polarity)*
X+	9	0	18
X–	3	6	0
A+	3	6	0
A–	9	0	18
AX+	9	18	0
AX–	3	0	6
Context+	3	0	6
Context–	9	18	0
ΔP_X	18/24 – 6/24 =.5	18/24 – 6/24 =.5	18/24 – 6/24 =.5
ΔP_A	12/24 – 12/24 = 0	24/24 – 0/24 = 1	0/24 – 24/24 = –1
$\Delta P_{X\|(no\ A)}$	9/12 – 3/12 = .5	0/6 – 0/18 = 0	18/18 – 6/6 = 0
$\Delta P_{X\|A}$	9/12 – 3/12 = .5	18/18 – 6/6 = 0	0/6 – 0/18 = 0

Note. The frequency of the presence (+) and absence (–) of the effect for each trial type is shown in the first eight rows. ΔP is the difference in the conditional probability of the effect given the presence and the absence of the target (e.g., $\Delta P_X = P(E|X) - P(E|\text{no X})$. The conditional probabilities for X in the presence and absence of A are calculated using the frequencies of the effect in the rows where A is or is not present. $\Delta P_{X|A}$ is the contingency of X conditional on the presence of A (i.e., calculated from only the trials when A is present; $\Delta P_{X|A} = P(E|X, A) - P(E|\text{no X}, A)$), and $\Delta P_{X|(no\ A)}$ is the contingency of X conditional on the absence of A; $\Delta P_{X|(no\ A)} = P(E|X, \text{no A}) - P(E|\text{no X}, \text{no A})$.

context there was no indicator so that all indicators were off on observations when both A and X were absent.

To study blocking and enhancement, each participant observed three contingency treatments. These were same polarity (.5/1) cue competition, opposite polarity (.5/–1) competition and a control contingency (.5/0). The first number in the treatment designation represents the contingency for the target X and this contingency was always moderately positive ($\Delta P_X = .5$). The second number represents the contingency for A that was either perfectly positive ($\Delta P_A = 1$), zero ($\Delta P_A = 0$), or perfectly negative ($\Delta P_A = -1$). The actual frequencies of the events and the frequency of trials they were paired with a life form are shown in Table 1.

The cause similarity treatments (the presence or absence of the common element) and the contingency treatments (.5/0, .5/1, and .5/–1) were present-

ed to all participants in a 2 × 3 within-participants design (i.e., there were six within-participants treatments). The two levels of the context salience treatment (i.e., the presence or the absence of a context indicator light on all trials) were each presented to half of the participants (i.e., it was a between-participants factor).

Method

Participants

Ninety-six McGill undergraduate students (67 (70%) female, mean age = 20.3 years, SEM = .135) were recruited from the McGill Psychology Participant Pool. Course credit was given for participating. All participants gave informed consent and were debriefed following the experiment.

Apparatus

Experiment 1 used "Alien Life I", a computerized causal reasoning task. It was presented to each participant in a quiet laboratory setting. Up to three participants were tested simultaneously, each seated in front of one of three iMac desktop computer stations separated by a low partition.

Procedure

The experimental procedure was approved by a McGill University Human Research Ethics Committee. Participants completed the experiment at their own pace. The task took approximately forty-five minutes to complete. Participants began by reading a set of instructions presented on the computer screen. They were instructed to imagine themselves as astronauts visiting six different planets. Their goal was to use the information about the planets' environments to determine whether or not alien life forms would be found on that planet. On the computer screen's virtual spaceship display, participants viewed five indicator lights labeled A-E, each of which signaled a different environmental variable. On each trial various combinations of indica-

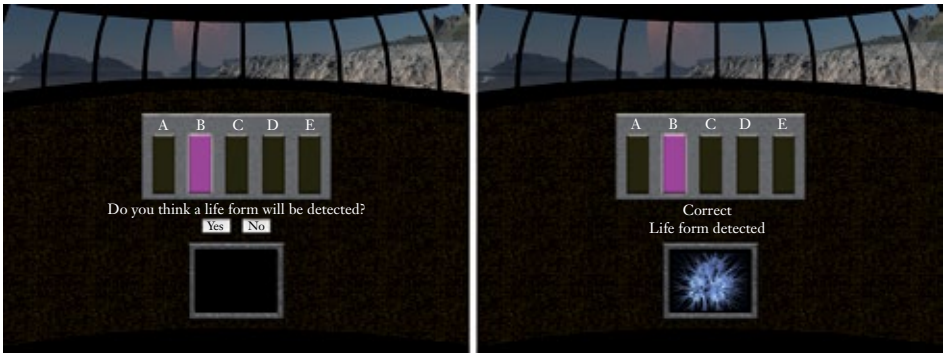

Figure 2. Left: Display panel showing five indicator lights, one of which is illuminated. The participant is asked to predict whether an alien life form will be detected. Right: Feedback screen that follows the participant's prediction. In this case, the participant correctly predicted the presence of an alien life form.

tor lights were illuminated. Light indicators represented each of the three cues (X, A, Context). If there were no common elements, a single indicator represented X and another represented A. If there was a common element, X and A were represented by two indicators; one was common to both A and X and the second was unique to that cue. If the context was salient another indicator was illuminated on all trials. The left panel of figure 2 shows a representation of the display panel.

The Salient Context Group (30 (64%) females, mean age = 20.3 years, SEM = .196) was exposed to the salient context and the Non-Salient Group (37 (76%) females, mean age = 20.3 years, SEM = .188) was not. The illuminated indicator lights for the different trial types are shown in Table 2. For the Salient Group one indicator light represented the salient context. This light was not illuminated for the Non-Salient Group. There were six treatments. The target cue (X) always had a moderate contingency with the outcome ($\Delta P_X = .5$). The competing cue (A) was either strongly positive ($\Delta P_A = 1$), strongly negative ($\Delta P_A = -1$) or zero/null ($\Delta P_A = 0$). All participants were exposed to the three types of contingency treatments (.5/0, .5/1, .5/–1) when X and A shared a common element (the Common Elements conditions) and when they had only their own unique elements (the No Common Elements conditions). Thus, each participant viewed six conditions: three for common elements and three for unique elements. Each of these six conditions was presented as a different planet in "Alien Life I".

TABLE 2. The presence and absence of indicator lights (elements) for the various stimulus configurations of Experiment 1.

Type of trial	*Context Indicator*	*A Indicator*	*X Indicator*	*Common Element Indicator*	
				Common Element Condition	*No Common Element Condition*
Salient Context Group					
A	+	+	–	+	–
X	+	–	+	+	–
AX	+	+	+	+	–
Context	+	–	–	–	–
Non-Salient Context Group					
A	–	+	–	+	–
X	–	–	+	+	–
AX	–	+	+	+	–
Context	–	–	–	–	–

Note. (+) indicator on, (–) indicator off.

There were forty-eight training trials for each planet (i.e., 48 trials per contingency treatment). To minimize order effects, the training trials within each treatment were randomly intermixed. The order of the six within-participant treatments was counterbalanced across participants as closely as possible. The assignment of the indicator lights to the cues was randomly determined. Each training trial showed a set of on/off indicators above the question: "Do you think a life form will be detected?" Participants answered by using the mouse to click either "yes" or "no". Following this feedback ("correct" or "incorrect") was provided. The presence of the outcome was an image of an alien form shown at the bottom of the screen (see right panel of figure 2) whereas the absence of the outcome was a blank box. Each of the six planets had different coloured light indicators, different background images and a different image for the alien life form found on that planet.

After observing all forty-eight trials of each contingency treatment, participants were asked to rate whether A, X, and the context signals the presence of an

alien life form. For the A and X ratings, participants were shown the indicator lights representing each of these causes while the context light was off. For the context ratings, only the context indicator was on for the Salient Context Group, whereas all indicators were off for the Non-Salient Context Group. Participants were asked the following question for each combination of indicator lights: "Please indicate what this combination of chemicals signals. +100 means that a life form will definitely be detected. 0 means that a life form is no more or less likely to be detected. –100 means that a life form will definitely NOT be detected. Please make your estimate using the scale below and click on the OK button". Data were analyzed with standard mixed or repeated measures analyses of variance (ANOVA) and t-tests. Effect size was analyzed using eta-squared (η_p^2) for the ANOVAs and Cohen's d_z for the t-tests (cf, Lakens, 2013).

Results

Ratings of X

The top panels of figure 3 shows participants' ratings of the causal relationship between the target X and the outcome. In both context conditions, the positive same-polarity competing cue A (Treatment .5/1) forced ratings of X well below zero. This is consistent with the strong contrast effect mentioned earlier. Moreover, when A was strongly negatively correlated with the outcome (Treatment .5/–1), ratings of X were enhanced above the control condition's (Treatment .5/0) moderately positive judgments of X. That is, compared to the control (.5/0) treatment, ratings of X were blocked in Treatment .5/1 and enhanced in Treatment .5/–1. The second clear finding is that the presence of a constant contextual cue moderated the magnitude of judgments of X in all treatments. That is, the absolute magnitude of judgments of all contingencies seemed to be reduced. Surprisingly perhaps, the presence of a common element shared by A and X, which should have increased generalization and thus moderated the blocking and enhancement effects, seemed to have very little effect.

Analyses of variance confirmed these impressions. First, a 2(3x2) mixed design analysis of variance (ANOVA) with context salience as the between-participants factor (Non-Salient, Salient) and contingency (.5/0, .5/1, .5/-1) and common elements (common elements, no common elements) as repeated factors was used to test the effects of each factor among all participants.

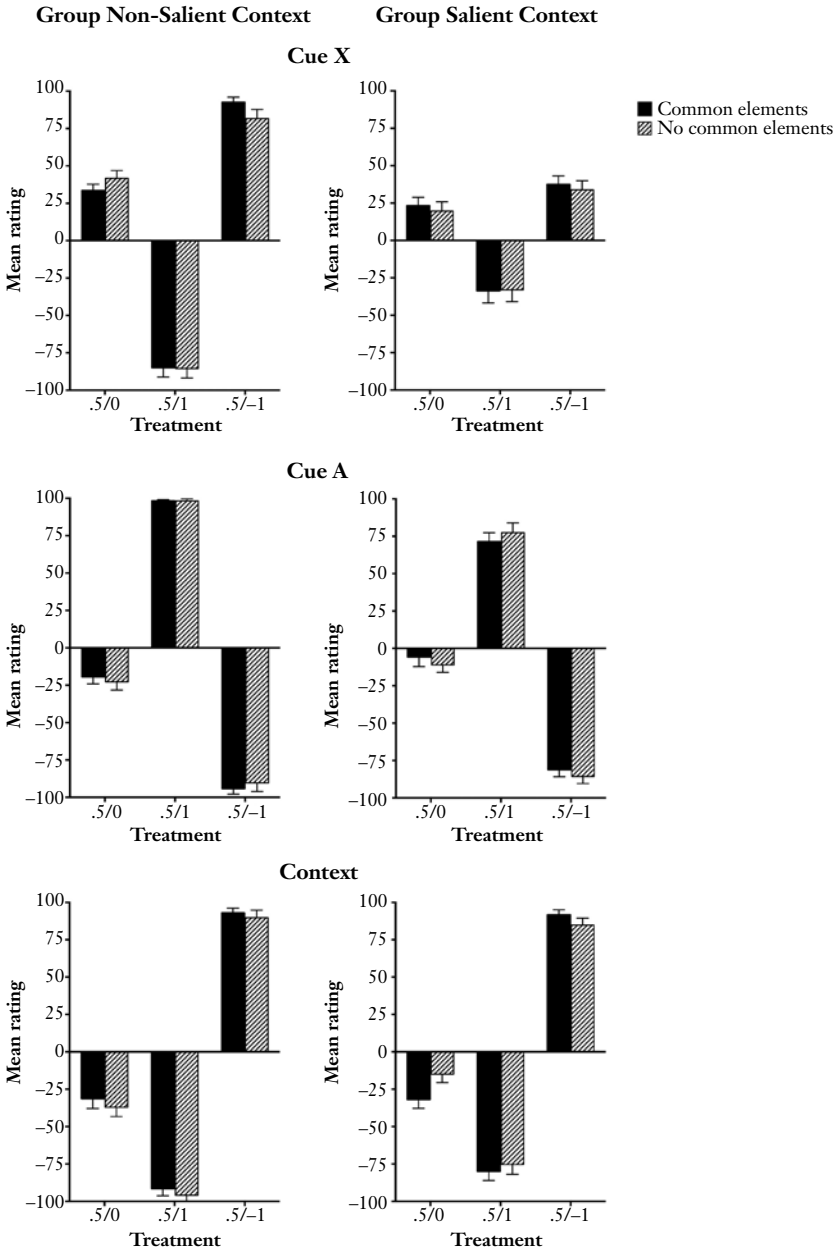

FIGURE 3. Mean causal ratings of cue X (top panels), cue A (middle panels), and the context (lower panels) in Group Non-Salient Context (left panels) and Group Salient Context (right panels) in Experiment 1. Error bars represent the standard error of the mean.

Judgments of X differed by contingency, $F(2,188) = 273.22$, $p < .001$, $\eta_p^2 = .744$. The main effect for the common element manipulation was not reliable, $F(1,94) = .30$, $p = .584$, $\eta_p^2 < .001$. Context salience did not change the pattern of ratings of X, but did attenuate them. This is reflected in a reliable interaction, $F(2,188) = 48.26$, $p < .001$, $\eta_p^2 = .339$.

As a consequence of this interaction we used repeated measure ANOVAs in each context group. Ratings of X in the Non-Salient Context group were analyzed with a two-factor 3 (.5/0, .5/1, .5/–1) by 2 (common elements, no common elements) ANOVA. The analysis confirmed that ratings of X were different in the three contingency treatments, $F(2,96) = 321.01$, $p < .001$, $\eta_p^2 = .870$, but that the common element manipulation had no reliable effect (no main effect nor any interaction with contingency; maximum $F(2,96) = 2.45$, $p = .092$, $\eta_p^2 = .048$). Two 2x2 ANOVAs compared each contingency treatment with the control. Compared to the control treatment .5/0, estimates of X were significantly blocked and significantly enhanced in the .5/1 and .5/–1 treatments, respectively, minimum $F(1,48) = 71.47$, $p < .001$, $\eta_p^2 = .598$.

The analysis of the ratings of X in the Salient Context group largely paralleled that of the Non-Salient Context group. In the ANOVA for contingency (.5/0, .5/1, .5/–1) and common elements, there was a main effect of contingency ($F(2,92) = 40.16$, $p < .001$, $\eta_p^2 = .465$) but no effect of the common element manipulation ($F(1,46) = .19$, $p = .664$, $\eta_p^2 = .004$), and no interaction, $F(2,92) = .09$, $p = .912$, $\eta_p^2 = .002$. The 2x2 repeated measure ANOVAs confirmed reliable blocking of X in Treatment .5/1 ($F(1,46) = 37.87$, $p < .001$, $\eta_p^2 = .452$), as well as a trend towards significance for enhancement in Treatment .5/–1, $F(1,46) = 3.79$, $p = .058$, $\eta_p^2 = .076$. As depicted in figure 3 and as expected from previous analyses, the common element manipulation did not have any effect on ratings of X in the Salient Context group, maximum $F(1,46) = .06$, $p = .802$, $\eta_p^2 = .002$.

Ratings of A

Ratings of A are shown in the middle panels of figure 3. When A was strongly positive or negative, mean ratings were very extreme in the Non-Salient Context group (absolute value of mean > 90). Again, the salient context moderated them somewhat (means ranged from 70 to 90). As has been found before, ratings of the zero contingency A were modestly negative (range –5 to

–30). Again, these were moderated by the salient context, as they were closer to zero in the Salient Context group.

The mixed design 2 (3 × 2) ANOVA confirmed these impressions. There was a reliable effect of contingency, $F(2,188) = 876.93$, $p < .001$, $\eta_p^2 = .903$, and the context moderation effect was confirmed by a reliable interaction, $F(2,188) = 11.59$, $p < .001$, $\eta_p^2 = .109$. No other effects reached significance, all other Fs < 1. As these results are straightforward we report no post hoc tests except to analyze whether the ratings of the zero contingency were pushed below zero. In the Non-Salient Context group both ratings (i.e., ratings of A in the .5/0 treatments) were below zero, minimum $t(48) = 4.00$, $p < .001$, $d_z = .342$. In the Salient Context group only the ratings in the no common elements condition were significantly less than zero, $t(46) = 2.21$, $p = .032$. $d_z = .323$ and $t(46) = .86$, $p = .393$, $d_z = .125$.

Ratings of the context

Ratings of the context are rarely reported and people's judgments of it are often inferred by its effect on other cues, but here we were able to assess people's impressions of the context directly. Context ratings are useful as they allow us to evaluate learning mechanisms that rely on the context to explain an effect. For example, a common explanation of the mechanism by which negative contingencies are represented relies on the notion that people form a strongly positive context representation (Baker, 1977; Rescorla & Wagner, 1972). Our participants' judgments of the context (lower panels of figure 3) are strongly consistent with this notion. In the treatments where A was negative (.5/–1), the context was judged strongly positive. When A was positive (.5/1 treatments), the context was judged negative. Moreover, when A was zero but X was modestly positive (.5/0 treatments), the context was judged to be modestly negative. Again, judgments appeared to be moderated by the context manipulation, with the Salient Context group giving stronger context ratings. Finally, the common element manipulation seemed to have little effect.

The 2(3 × 2) mixed design ANOVA again supported these claims. Again the main effect for contingency was reliable, $F(2,188) = 841.20$, $p < .001$, $\eta_p^2 = .899$, and none of the other main effects or interactions was reliable, maximum $F(1,94) = 3.72$, $p = .057$, $\eta_p^2 = .038$. As with ratings of X, we carried out analyses comparing the control .5/0 contingencies with the same (.5/1)

and opposite (.5/–1) polarity treatments and found that each of these comparisons was reliable, minimum $F(1,46)= 79.95$, $p < .001$, $\eta_p^2 = .654$. As well we compared judgments of the context in the control (.5/0) contingencies with zero and found each of these to be reliable, confirming that these were all less than zero as was expected by the contrast hypothesis (Darredeau et al., 2009), minimum $t(46) = -2.71$, $p = .01$. These findings are again consistent with the contrast hypothesis, whereby ratings of the context are "pushed" away from the stronger competing causes (X in Treatment .5/0 and A in treatments .5/1 and .5/–1).

The results of Experiment 1 are quite clear. Consistent with the contrast hypothesis put forward by Darredeau et al. (2009) we found blocking of X in the same polarity (.5/1) treatments and enhancement in the opposite polarity (.5/–1) treatments. Moreover, we found blocking past zero in all conditions including blocking of A in the .5/0 control condition and in context judgments. That is, in the .5/1 treatments the stronger cue A pushed ratings of X below zero, and in the .5/0 treatments the stronger cue X pushed ratings of A and the context below zero. Furthermore, the presence of a salient context attenuated judgments in nearly all cases. However, a more surprising result was that the inclusion of common elements seemed to have little effect on judgments.

One potential reason why the inclusion of a common element may not have been effective is because the common element used here had very much in common with the standard discrete elements used in conditioning. The basic tenets of stimulus sampling theory, however, posit that a cue is composed of multiple elements and not symbol-like single representations (Atkinson & Estes, 1963). It is thus possible that a more complex stimulus array might encourage participants to discriminate the common element treatment from that containing only unique elements.

Experiments 2A and 2B

Experiments 2A and 2B used that same basic design as Experiment 1 but used a more complex display that is more analogous to the basic assumptions of stimulus sampling theory. Unlike Experiment 1, the context manipulation was carried out separately in these experiments. Experiment 2A included a nonsalient context and was conducted first, and Experiment 2B included a salient context and was conducted later.

Method

Participants

Experiment 2A was carried out first and analyzed before carrying out Experiment 2B. Sixty-two McGill undergraduate students (42 (71%) females, mean age = 21.8 years, SEM = .583) enrolled in a Psychology course (Animal Learning and Theory, PSYC 301) participated. Subsequently, ninety-one McGill Undergraduate students were recruited from the McGill University Psychology Participant Pool to participate in Experiment 2B (68 (76%) females, mean age = 20.4 years, SEM = .212). Course credit was given for participation.

Procedure

The design of the two experiments was the same as Experiment 1. Participants predicted the presence of alternative life forms on six planets. There were three contingencies between A, X and the outcome (.5/0, .5/1. and .5/–1), and each contingency was presented in the absence and the presence of common elements. In Experiment 2A there was no contextual cue present on each trial (similar to the Non-Salient Context group in Experiment 1), whereas in Experiment 2B a set of contextual cues were present on each trial (analogous to the Salient Context manipulation in Experiment 1). As in Experiment 1, there were forty-eight trials of each contingency and similar counterbalancing was carried out.

The display shown in figure 4 consisted of 0, 3, 8 or 11 elements depending on whether the context was present or not. The salient context consisted of three elements that were present on all trials (in Experiment 2B) but that were absent in the non-salient context manipulation (Experiment 2A). Thus, the context-alone trials consisted of either 0 or 3 elements. Each cue A or X consisted of 4 elements. In the no common elements treatments, all 4 elements of each cue were unique (i.e., different from those of the other cue). However, in the common elements conditions, the cues A and X shared 2 elements and had 2 unique ones. To model the notion that the sensory buffer for a cue is of a constant size, any time a discrete cue or compound was presented, 8 elements were active in the array. The 3 context elements were also illuminated if appropriate making up to 11 elements. If it were an all-unique

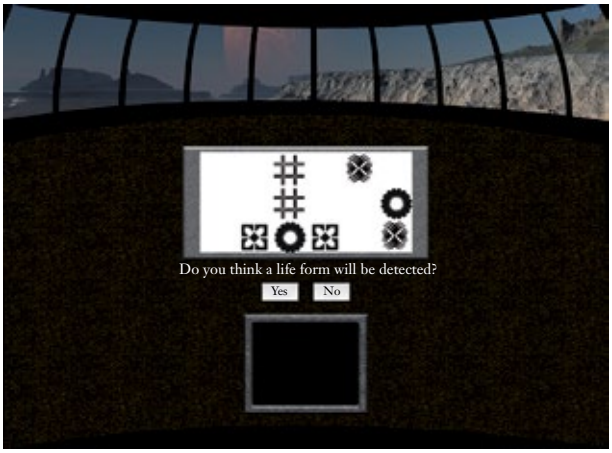

FIGURE 4. Example of a display panel used in Experiments 2A and 2B. Each symbol represents an element that could belong to cues A or X or the context.

elements AX compound trial, the 8 unique elements representing those cues would be present. For a common-elements compound trial the same rule applied but because the cues had common elements, each common element appeared twice in the array. Finally, for a single cue presentation, each of its 4 elements would appear twice. The array was a 6 column by 3 row matrix and on any trial the elements could appear in any position.

After participants viewed all 48 training trials in a given treatment, they rated the efficacy of A, X, and the context as they did in Experiment 1. For the A and X ratings, participants were shown only the 8 symbols representing each of these causes in the absence of the 3 context symbols. For the context ratings, only the 3 context symbols were shown in Experiment 2B, whereas the symbol display was blank (i.e., none of the symbols appeared) in Experiment 2A.

Results

Ratings of X

The mean estimates of cue X for Experiments 2A and 2B are shown in the top panels of figure 5. It is clear from this figure that we have systematically

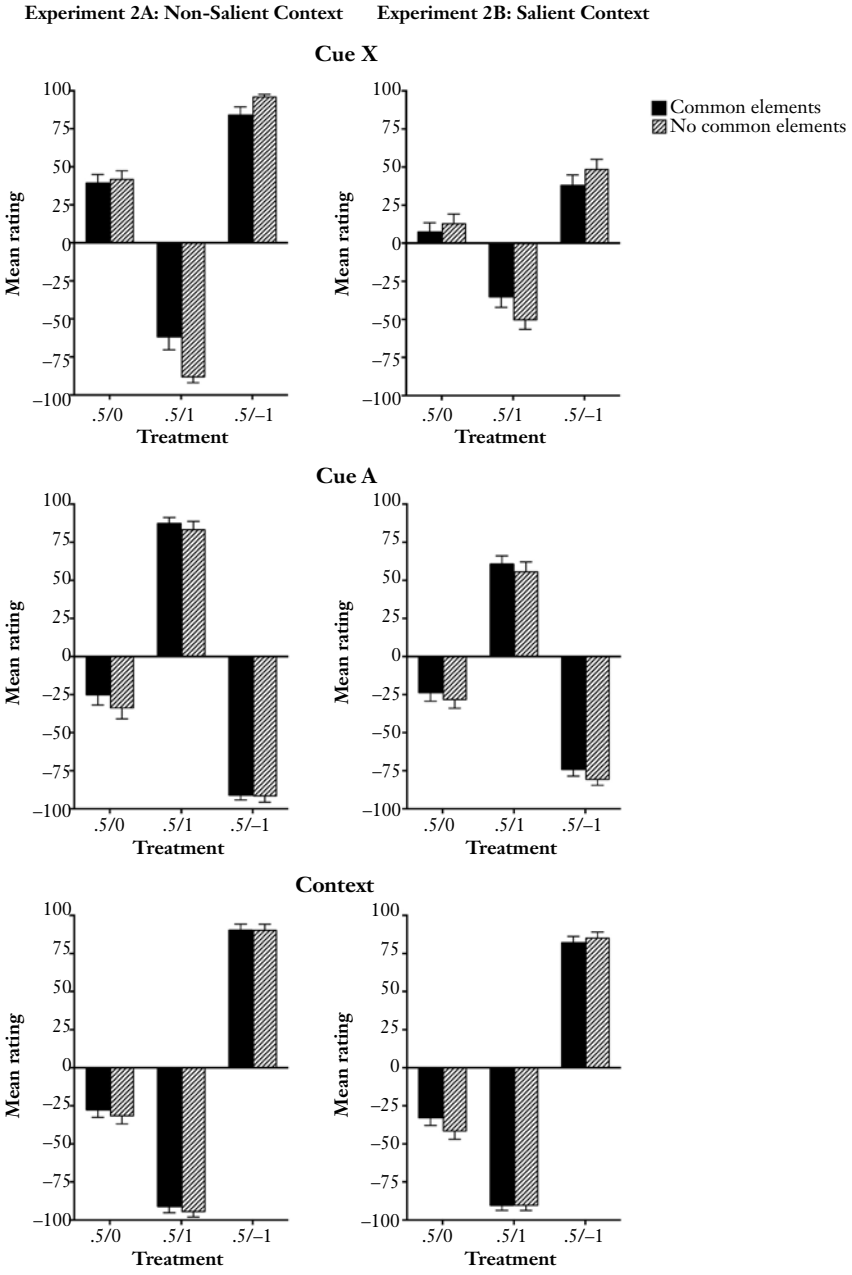

FIGURE 5. Mean causal ratings of cue X (top panels), cue A (middle panels), and the context (lower panels) in Experiments 2A (left panels) and 2B (right panels). Error bars represent the standard error of the mean.

replicated the results of Experiment 1. Judgments of X are pushed below zero in the same polarity (.5/1) treatments (i.e., X was "blocked" past zero), and are enhanced above the control (.5/0) in the opposite polarity treatments (.5/–1). Moreover, although this is a between experiment comparison and hence confounded by time of year and having large but different sample sizes, it seems that the presence of a salient context (Experiment 2B) moderates judgments of all contingencies just as it did in Experiment 1. However, the novel finding of this experiment is that the presence of common elements appears to moderate judgments of X in a way parallel to the effects of the salient context and in a manner consistent with that expected from the common elements interpretation of stimulus generalization theory.

We do not report a full analysis comparing judgments in Experiments 2A and 2B because they are very similar but independent experiments. Nevertheless, as noted above, comparing Experiments 2A and 2B represents a systematic replication of the ability of the salient context to attenuate judgments of X. So to give some indication of the strength and reliability of this replicated effect we report that the interaction consistent with moderation by the context was reliable, $F(1,151) = 34.42$, $p < .001$, $\eta_p^2 = .186$, as was the contingency by common elements interaction that is consistent with the moderation by the common elements manipulation, $F(2,151) = 9.91$, $p < .001$, $\eta_p^2 = .062$. Moreover, one way ANOVAs comparing the absolute magnitude of estimates between experiments (e.g., comparing .5/0 with common elements and with or without a salient context) were all reliable, minimum $F(1,151) = 5.63$, $p = .019$, $\eta_p^2 = .036$.

We analyzed the two experiments separately to investigate the separate effects of contingency and the presence of common elements. In Experiment 2A the main effect for contingency was reliable, $F(2,122) = 352.87$, $p < .001$, $\eta_p^2 = .853$, but that for common elements was not, $F(1,61) = .96$, $p = .331$, $\eta_p^2 = .015$. As in Experiment 1 with the context, this was because the elements manipulation increased judgments in Treatment .5/–1 and decreased them in Treatment .5/1, thus leaving the means relatively unaffected. However, as in Experiment 1, the interaction that represents this effect was reliable, $F(2,122) = 7.92$, $p < .001$, $\eta_p^2 = .115$. Comparisons of the common elements and no common elements contingencies confirmed that enhancement and blocking were attenuated by the presence of common elements, minimum $t(61) = 2.24$ $p = .029$, $d_z = .285$.

Although the absolute values of the means are much smaller in Experiment 2B, the pattern of results is similar. There was a reliable contingency

effect, $F(2, 180) = 75.18$, $p < .001$, $\eta_p^2 = .455$. Again the interaction that reflects moderation by common elements was reliable ($F(2,180) = 3.43$, $p = .034$, $\eta_p^2 = .036$), but there was no main effect for common elements, $F(1,90) = .00$, $p = .956$, $\eta_p^2 = < .001$. We carried out 2x2 ANOVAs comparing the control with the other treatments (i.e., .5/1 v .5/0 and .5/–1 v .5/0). The main effects for contingency were reliable, minimum $F(1,90) = 29.97$, $p < .001$, $\eta_p^2 = .250$, indicating that each of these contingencies differed from one another (i.e., the blocking and enhancement effects were reliable). Because the interactions were not reliable we carried out a third 2x2 ANOVA comparing .5/1 and .5/-1, the two treatments in which the moderation by common elements was expected. In this ANOVA the interaction was reliable, $F(1,90) = 4.80$, $p = .031$, $\eta_p^2 = .051$, indicating that the common element treatment did moderate the judgments of X. This effect was quite weak possibly because the salient context had already reduced the difference between judgments of X.

Ratings of A

The results for cue A were fairly straightforward. Participants discriminated between the positive, negative, and zero contingencies for A (see middle panels of figure 5). The presence of a salient context moderated this effect although this mechanism was much more obvious with the non-zero contingencies. The presence or absence of common elements had little effect on A judgments. Finally the judgments of the zero contingency (in the .5/0 treatments) which had been contrasted with the modest positive X contingency were negative. The first two impressions were confirmed by the omnibus ANOVA comparing the two experiments. The main effect for contingency was reliable as was the contingency by experiment interaction, minimum $F(2,151) = 10.44$, $p < .001$, $\eta_p^2 = .064$. The main effect for common elements was nearly reliable, $F(1,151) = 3.44$, $p = .066$, $\eta_p^2 = .022$.

As with X, we carried out individual 2x3 common element by contingency ANOVAs on the two experiments. In both Experiments the only reliable effect was the main effect for contingency, minimum $F(2, 180) = 216.08$, $p < .001$, $\eta_p^2 = .706$; maximum non-reliable $F(1,90) = 2.54$, $p = .115$, $\eta_p^2 = .027$. To confirm that the three contingencies differed from one another and that there was no reliable effect of the common elements we carried out four 2x2 common element by contingency ANOVAs comparing the control.5/0 treatment

with the same (.5/1) and opposite (.5/–1) polarity treatments in each of Experiments 2A and 2B. Only the main effects for contingency were reliable, minimum $F(1, 90) = 87.20$, $p < .001$, $\eta_p^2 = .492$.; maximum non-reliable $F(1,90) = 2.09$, $p = .152$, $\eta_p^2 = .023$. Finally, all estimates of A in the control .5/0 treatment were reliably less than zero, which is again consistent with the contrast arguments that the modest X ($\Delta P = .5$) pushes these estimates past the zero point, minimum $t(62) = 3.70$, $p < .001$, $d_z = .465$.

Ratings of the context

The ratings of the context for Experiments 2A and 2B are shown in the lower panels of figure 5. This figure shows that neither the common elements nor the context saliency manipulation influenced the context ratings, but the different A and X contingencies did. This conclusion is supported by the omnibus 3x2x2 ANOVA comparing the two experiments. The only reliable effect was that of contingency, $F(2,302) = 1098.57$, $p = .001$, $\eta_p^2 = .879$, maximum non-reliable $F(2,302) = 1.07$ $p = .345$, $\eta_p^2 = .007$. As with A, the estimates of the context were all below zero in the control .5/0 contingencies, minimum $t(62) = 5.44$, $p < .001$. $d_z = .685$.

Discussion

The results we have reported are clear. We have replicated earlier findings that the presence of strong competing cues alters judgments of weaker cues (Baker et al., 2000; Darredeau et al., 2009; Vallée-Tourangeau et al., 1998). Thus, people are not just sensitive to the simple contingency between a potential cause and effect, but somehow integrate the information about other potential causes of the same effect. One explanation of this is that people are sensitive to the relative information provided by each cue and reason that, because the strong cue is perfectly informative of the outcome, there is no rational reason to ascribe any causal power to the weaker cues. Formally, in terms used by Cheng and her colleagues (Cheng & Novick, 1992), the conditional probabilistic contrast for the weaker cause given the presence or the absence of the stronger one is zero (i.e., $\Delta P_{X|A} = \Delta P_{X|(no\ A)} = 0$). The weaker cause provides no information above and beyond the information provided

by the stronger cause, so the weaker cause is ascribed no power. However, the fundamental results of these experiments are not consistent with this view. Indeed, in the cross polarity contingency (.5/–1) treatments, the strong negative cue A was entirely informative of the outcome generating a contrast for the target X of zero, yet, rather than weakening judgments of X, it enhanced them. This enhancement is inconsistent with the notion of conditional contrasts (Cheng & Novick, 1990; 1992). Furthermore, it is unlikely to be a consequence of the strong cue's negative valence because other experiments by Darredeau and her colleagues have found similar negative enhancement when the strong cue was positive and the moderate target negative (Darredeau et al., 2009).

The effects of the context and common elements manipulations

The effect of the common elements in Experiments 2A and 2B can be explained by considering cue similarity or generalization. Adding common elements to two cues increases their similarity (Atkinson & Estes, 1963; Pearce, 1987). More similar cues are treated as more alike and impressions of one cue are more likely to generalize to the other. Thus, adding common elements should reduce differences between a participant's impressions and hence ratings of two cues because it would effectively average the ratings. Increasing the salience of the context also increases similarity of the two cues because it effectively adds more common elements. But beyond that, the "extra" context elements are present on all trials and thus are paired with all outcomes. This allows them to share in more of the combined associative strength that is available, thus reducing the perceived strength of cues A and X.

Thus, in the present experiments one would expect the presence of a strong opposite polarity cue to push judgments of the target X away from the position of the strong competitor A. However, to the extent they are similar, the generalization between the strong impression of A and the moderate impression of X should moderate the size of the cue interaction effects. Moreover, to the extent that the strong context controls more of the outcome expectancy, this should also moderate differences between A and X. And this is just what we found: adding common elements and increasing the context salience moderates both blocking and enhancement.

One possible alternative explanation for our results is that people simply report the proportion of outcomes in the presence of X, A, and the context on trials in which each occurs in the absence of the other cues (these predictions can be derived from the PCM as well if one assumes that the cue ratings are based on the probability of the effect given the cue and the context). This account suggests that people only considered the events on cue-alone trials independent of what happened on the other trials. At first glance this account seems compelling because the ordinal predictions of X in the three treatments follows this metric. Moreover, it could be extended to explain how removing the salient context in the salient context treatments reduced estimates of A and X compared to the non-salient context treatments. On the context test trials and on all test trials in the non-salient context treatments, the participants were presented with the exact stimulus configuration they had seen in training, whereas this was not the case for the A and X test trials in the salient context treatments. On these latter test trials, participants had never seen A and X in the absence of the salient context cue. Thus through generalization decrement or lack of confidence that the A and X alone were the same stimuli they had already seen, their judgments were reduced. Indeed, because of the strong judgments of the context alone, it could be argued that the strong "contrast" effect on X was entirely controlled by the participants' impression of the context's strength. The smaller "contrast" to X when the salient context cue is dropped might be controlled by the residual context elements that were not removed (i.e., the background). Thus, the idea that the ratings of A, X and the context are based on the P(E|A alone), P(E|X alone), and P(E|Context alone), respectively, and that there is some generalization decrement for the ratings of A and X in the salient context treatments could explain all of our results. This argument, of course, requires the assumption that the participants perceive the –100 to +100 rating scale, in which participants are told that negative numbers represent a decrease in outcome probability from baseline, is instead interpreted as a 0 to 1 probability scale.

The above argument, however, relies on the idea that participants base their judgment of a cue only on trials with that cue alone. But Darredeau et al. (2009, pp. 8 and 9) studied contingencies where strong cues that were paired with varying numbers of the total number of outcomes provide arguments against this claim. They report differences between ratings of a target cue X in treatments with identical P(E|X alone) (e.g., compare ratings of X in Treatments .5/.67 0 and .5/ 1 0 in Darredeau et al.'s figure 1). Moreover, some-

times people report blocking rather than enhancement in the .5/–1 treatments (e.g., Baker et al., 1993) and this is the exact opposite of this mechanism (these findings are also inconsistent with a contrast mechanism). Thus, this simple explanation cannot reasonably be considered to be a general explanation of what happens in these experiments although it cannot be eliminated entirely here. White (e.g., 2005) has offered more complex models such as his evidence evaluation model that do consider interaction between cues. But they would have trouble explaining the context effects, the results of Darredeau et al (2009), and any other case in which the same objective contingency generates different judgments.

We have simulated these experiments using the Rescorla-Wagner model and, although it does not provide a complete account of our findings, it does predict that the salience of the context would have a large influence on the blocking and enhancement effects, both pre-asymptotically and at asymptote. In contrast to the large and stable context effect, it predicted only a moderate effect of common elements, and this only pre-asymptotically. This large difference between the two effects is consistent with the data, as the common elements had a much more modest influence on causal ratings, which was significant only in Experiments 2A and 2B. The model predicts that a salient context will reduce the blocking effect, and that the enhancement effect will be reversed (i.e., that blocking will occur in Treatment .5/–1 compared to Treatment .5/0). Although the first prediction was confirmed by our data, the second only received partial support: a salient context resulted in a weaker enhancement effect rather than a complete reversal of this effect. The simulations with the Rescorla-Wagner model provide an alternative mechanism that might explain the apparent contrast effects we observed without resorting to a contrast mechanism whereby causes are directly compared to one another. It is also possible to explain why the ratings of X in the blocking and enhancement treatments were never greater in magnitude than ratings of the context. A critical difference between our salient and non-salient context manipulations is that whereas in the non-salient context conditions the entire context (the background) is presented along with the target cue on test trials, this is not the case for the salient context conditions in which the target cue is presented with the background context, but the context indicator is turned off, so when a cue is tested, it is tested in the presence of only some of the context elements it was trained with. To simulate this possibility, we ran a second set of simulations in which there was a second context element that was present-

ed during training and the test. In this case the context elements (the background and the context indicator, if present) would acquire associative strength in both context conditions. The participants' ratings are based on the sum of the associative strengths of the cue and the background present at test. In the non-salient context condition, this means the sum of the associative strengths of the cue and the background context. In the salient-context condition, on the other hand, the background and the context indicator compete for associative strength during training, and only the background is presented along with the cue on test trials. This second set of simulations not only resulted in similar salient context and common elements effects on the ratings of X as those described above, but it also generated a pattern of associative strengths for the context consistent with the data.

So it is possible that the apparent contrast effects we observed were not the result of a contrast mechanism. Instead, such effects could be due to the ratings of X being influenced by the perceived strength of the context. This explanation however, has less success in accounting for the ratings of A. In Treatment .5/–1, the context acquires positive associative strength and removing some of its influence at test should cause the ratings of A to be more negative in the salient context condition, but we observed the opposite pattern.

Information processing accounts such as the probabilistic contrast model or conditionalization provide an indirect explanation of the moderation of judgements caused by the addition of common elements (Cheng & Novick, 1990, 1992, Spellman, 1996). According to these accounts, when multiple causes are present, judgments of a single cause are made by assessing what happens with that target cause when all of the other causes are held constant. Because the common elements are always present when A and X are present, a contrast involving the presence and absence for the common elements cannot be computed. The same could be argued for the context because it is always present. When one considers how many cues in the real world must pose this problem, it questions the generality of the theory. It is nonetheless possible that when participants are faced with this dilemma they decide that the common elements are weak candidates compared to the unique elements. Thus, when a decision is made about the compound of unique and common elements, that make up A or X, the presence of the common elements weakens the perceived strength of these cues and thus moderates judgments. This is a seemingly plausible explanation of stimulus generalization but it must be realized that it is an *a posteriori* ancillary assumption that does not arise di-

rectly from the *a priori* computations, or principles, of the model. One could just as easily assume, *a posteriori*, that elements that cannot be computed are ignored, which is in harmony with attention theories (Mackintosh, 1975), or that they take on the properties of cues with which they are regularly paired — the unique elements — or that the reasoner could assume any one of the interaction or other higher order contrasts that these theories allow and that, likewise, cannot be eliminated or substantiated. Moreover, this begs the question of what to do with the multiple unique elements; they occur with one another so no one contrast can be computed. If they are treated as a single compound or token, then this is not a problem. But then one needs a principle to explain which elements are to be agglomerated and which to be separated for independent analysis.

Blocking past zero

Another important finding that is not consistent with the cognitive information reduction argument (Baker et al., 1996, Spellman, 1996) is the fact that judgments of A are systematically below zero in the .5/0 control treatments. A correlated informative cue (X) can reduce the value of a weaker cue (A) to zero, but does not change its polarity. And the value of A is already zero. However, if X influences A through a contrast mechanism, then it could push the value of A into the negative range. While it might be argued that these negative estimates represent a scaling error (e.g., participants might consider –20 on the rating scale to be psychological zero), this argument is challenged by other data. Darredeau and her colleagues found that in a –.5/0 treatment X increased judgments of A above zero whereas in a .5/0 treatment, in a manner similar to the present experiment, X reduced judgments of A below zero (Darredeau et al., 2009). This crossover of the ratings of the zero A contingency is difficult to reconcile with the scaling error argument.

Furthermore, in both experiments the moderate X contingency in the control .5/0 treatments generated negative judgments of both the salient and the non-salient context. In Darredeau et al. (2009) we reported a crossover of the ratings of X in the same-polarity treatments, so the ratings of X in a –.5/-1 treatment were positive and the ratings of X in a .5/1 treatment were negative. Thus, a blocked negative contingency was rated more positively than a blocked positive contingency. We have called this effect "blocking past zero"

because a complete crossover of the ratings of the moderate positive and negative contingencies indicates that X was blocked past the participants' subjective zero. This supports the notion of contrast according to which the ratings of the moderate cue X are pushed away from the perceived causal efficacy of the stronger cue A. This contrast mechanism predicts that the perceived causal efficacy of X can be pushed below zero even though its objective contingency is positive.

The Rescorla-Wagner model predicts both crossover effects reported in Darredeau et al. (2009) when the context salience is low. It predicts a zero associative strength for X in Treatment .5/1 and a positive associative strength for X in Treatment –.5/–1, thus a moderate negative cue acquires more excitatory strength than a moderate positive cue. Similarly, the model predicts a small positive associative strength for A in Treatment .5/0 and a larger positive associative strength in Treatment –.5/0. However, the model does not predict a negative associative strength for X in Treatment .5/1, nor does it predict a negative associative strength for cue A in Treatment .5/0. So if we assume that the associative strength of a cue matches the participants' internal rating scale, then the model does not predict the negative ratings of X that we observed in Treatment .5/1, nor the negative ratings for A that we observed in Treatment .5/0.

It is possible, however, that the participants' subjective perception of causal efficacy did not match the rating scale they were asked to use when making their causal judgments. In that case, only the order of the ratings' magnitude can be interpreted, whereas the absolute ratings that were given should be interpreted with caution. If so, the Rescorla-Wagner model's predictions regarding blocking past zero are consistent with the present and Darredeau et al.'s findings, as its ordinal predictions match the order of the participants' ratings. But this calls into question the blocking past zero effects presented here because we compared the participants' ratings to an absolute value of zero, which may not reflect the participants' true subjective zero. It is worth noting, however, that Darredeau et al. (2009) provide a clear demonstration of blocking past zero effects that do not suffer from potential confounds regarding the scale interpretation. In their experiments, participants judged a blocked positive contingency as being more negative than a blocked negative contingency. This crossover of the ratings of contingencies of opposite polarity demonstrates blocking past the participants' subjective zero, regardless of where on the rating scale that subjective zero might be. Thus it seems that

this phenomenon is indeed robust and not an artifact resulting from the way participants interpret the rating scale. Blocking past zero effects are accounted for by the Rescorla-Wagner model (at least its ordinal predictions are consistent with the causal ratings), whereas statistical information processing accounts such as the Probabilistic Contrast Model (PCM) and conditionalization do not account for them.

Conclusion

We argue that a simple bottom-up automatic associative mechanism makes predictions that are generally, though not always, consistent with our findings. The model predicts a large context effect on blocking and enhancement, and a very modest pre-asymptotic effect of common elements. Although we did not demonstrate blocking in Treatment .5/–1 when the context was salient, we did find a reduction in the enhancement effect that was predicted by the model. Perhaps the context manipulation does indeed, at least partially, explain some of the inconsistent findings reported previously, as blocking and enhancement effects have both been reported with Treatment .5/–1. So the model could be a useful tool to further investigate the origin of these inconsistent findings. Finally, its ordinal predictions are also consistent with blocking past zero. However it should also be pointed out that the assumption that people report the probabilities of the outcome experienced on the single cue trials provides another parsimonious explanation of the present results. Nevertheless, as argued above, it does not do so well with the results of other similar experiments.

This is a step forward in providing a potential explanation for apparent contrast effects and we argue that associative models hold more promise than statistical models for modeling the conditions that foster enhancement rather than blocking. Regardless of whether there are fewer or more common elements, there is still a large proportion of unique elements to make judgments of the information provided by the cues, yet people seem to act in a manner consistent with an automatic associative process. Although it can always be argued that the presence of common elements or a more salient context steals attention away from, or in some other way interferes with, the appraisal of statistical information, this still does not account for enhancement rather than blocking in the cross polarity treatments.

References

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In Luce R. D., Galanter E., Bush R. R. (Eds.), *Handbook of mathematical psychology* (pp. 121-268). New York: Wiley.

Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology. Animal Behavior Processes*, *35*, 153-168.

Baetu, I., & Baker, A. G. (2012). Are preventive and generative causal reasoning symmetrical? Extinction and competition. *Quarterly Journal of Experimental Psychology*, *65*, 1675-1698.

Baetu, I., & Baker, A. G. (in press). Human learning about causation. In R. A. Murphy, & R. Honey (Eds.), *The Wiley-Blackwell handbook on the cognitive neuroscience of learning*. Hoboken, NJ: Wiley-Blackwell.

Baker, A. G. (1976). Learned irrelevance and learned helplessness: Rats learn that stimuli, reinforcers and responses are uncorrelated. *Journal of Experimental Psychology: Animal Behaviour Processes*, *2*, 131-141.

Baker, A. G., Berbrier, M. W., & Vallée-Tourangeau, F. (1989). Judgments of a 2×2 contingency table: Sequential processing and the learning curve. *Quarterly Journal of Experimental Psychology B*, *41*, 65-97.

Baker, A. G., & Mackintosh, N. J. (1977). Excitatory and inhibitory conditioning following uncorrelated presentations of CS and UCS. *Animal Learning and Behaviour*, *5*, 315-319.109

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, P., & Pan, M. (1993). Selective associations and causality judgements: Presence of a strong causal factor may reduce judgements of a weaker one. *Journal of Experimental Psychology*, *19*, 414-432.

Baker, A. G., Murphy, R. A., & Vallée-Tourangeau, F. (1996). Associative and normative models of causal induction: Reacting to versus understanding cause. In D. L. Medin, D. Shanks & K. Holyoak (Eds.). *The psychology of learning and motivation*, vol. *34* (pp. 1-45). San Diego CA: Academic Press.

Baker, A. G., Vallée-Tourangeau, F., & Murphy, R. A. (2000). Asymptotic judgment of cause in a relative validity paradigm. *Memory and Cognition*, *28*, 466-479.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545-567.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365-382.

Darredeau, C., Baetu, I., Baker, A. G., & Murphy, R. A. (2009). Competition between multiple causes of a single outcome in causal reasoning. *Journal of Experimental Psychology. Animal Behavior Processes*, *35*, 1-14.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology A*, *36*, 29-50.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, *4*, 123-124.

Hume, D. (1740). *A treatise of human nature* (1967 edition). Oxford: Oxford University Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.

Mackintosh, N. J. (1975). A Theory of Attention: Variations in the Associability of Stimuli with Reinforcement, *Psychological Review*, *82*, 276-298.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral Brain Sciences*, *32*, 183-198.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*, 61-73.

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

Spellman, B. A. (1996). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 34, pp. 167-206). San Diego, CA: Academic Press.

Vallée-Tourangeau, F., Murphy, R. A., & Baker, A. G. (1998). Causal induction in the presence of a perfect negative cue: Contrasting predictions from associative and statistical models. *Quarterly Journal of Experimental Psychology B*, *51*, 173-191.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol 34. Causal learning* (pp. 47-88). San Diego, CA: Academic Press.

Wasserman, E. A., Elek, S. M., Chatlosh, D. J., & Baker, A.G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 174-188.

White, P. A. (2005). Cue interaction effects in causal judgment: An interpretation in terms of the evidential evaluation model. *Quarterly Journal of Experimental Psychology B*, *58*, 99-140.

Alleviation of Acute Caffeine Withdrawal Reinforces Flavor Liking

Paula J. Durlach
U. S. Army Research Lab - Human Research and Engineering Directorate*

Abstract. Research on the ability of caffeine to affect liking of flavors with which it has been paired is reviewed. Several experiments have shown that repeated consumption of a novel-flavored beverage paired with caffeine can result in increased rated liking for the beverage; however, this is only the case in self-selected habitual caffeine consumers who are acutely caffeine deprived both at the time of flavor-caffeine pairings, and at the time of flavor rating. Enhanced liking fails to occur if self-selected habitual caffeine consumers have been chronically withdrawn from caffeine prior to flavor-caffeine pairings. Nor is it acquired or expressed if acutely caffeine deprived users are given a caffeine preload prior to pairings or ratings. The sensitivity to deprivation state of enhanced liking based on flavor-caffeine pairings suggests that it is due to the ability of the caffeine to alleviate withdrawal symptoms, rather than to any state-independent positive effects of caffeine. Research participants who were not given caffeine in their target beverage were just as likely to say that the beverage contained caffeine as those that did receive it in their target beverage, suggesting that this phenomenon is an example of evaluative conditioning without contingency awareness.

In this chapter research on the ability of caffeine to affect liking of flavors with which it has been paired is reviewed. Caffeine is the most widely used pharmacologically active substance in the world (Juliano & Griffiths, 2004). According to Heckman, et al. (2010), coffee and tea are the two most prevalent sources, although soft drinks and energy drinks are also common. A recent survey by Mitchell, et al. (2014) estimated that mean daily caffeine in-

* Postdoctoral Fellow under N. J. Mackintosh, University of Cambridge (1983-1987).

take in the U. S. is 165 mg/day or 2.2 mg/kg of body weight; but they also noted the difficulty of estimating this amount, due to the wide variation in the caffeine content of brewed coffee and tea. When trying to estimate average caffeine consumption, researchers have frequently adopted the convention of 125 mg/cup of brewed coffee, 70 mg/cup of instant coffee, and 60 mg/cup of tea, following the amounts proposed by James (1991).

Caffeine produces a range of physiological effects. Subjective psychological effects include increased alertness, energy, and sociability; however, high doses may produce tension and restlessness (Juliano & Griffiths, 2004). Juliano and Griffiths (2004) identified headache, fatigue, irritability, and decreased alertness, concentration, and mood as among the effects of caffeine abstention by regular users of as little as 100 mg/day. These symptoms typically appear 12 to 24 hours after abrupt caffeine abstinence (Juliano & Giffiths, 2004). It has been suggested that the main reason for regular caffeine consumption is to forestall such negative withdrawal effects (e.g., Heatherley, Hancock, & Rogers, 2006; Rogers, 2000; Rogers, 2014; Schuh & Griffiths, 1997; Tinely, Yeomans, & Durlach, 2003), rather than because of any state-independent positive effects. The sensitivity of measurable caffeine effects to deprivation state has been found to vary with different measures, however (e.g., Addicott & Laurienti, 2009).

Several experiments have shown that repeated consumption of a novel-flavored beverage paired with caffeine can result in increased rated liking for the beverage; however, this is only the case in self-selected habitual caffeine consumers. Repeated pairings of a novel-flavored beverage with caffeine in self-selected habitual non- or low-consumers produces either no change or a decrease in liking (e.g., Dack & Reed, 2009; Richardson, Rogers, & Elliman, 1996; Tinley, Durlach, & Yeomans, 2004). In one of the earliest experiments, Richardson, Rogers, and Elliman (1996), using a double-blind procedure, gave university students a novel juice to drink with a capsule containing caffeine (100 mg) or placebo (corn flour), post-lunch, on ten occasions. Participants had been instructed to eat their normal lunch, but drink only water during, and two hours following lunch. The target juice was personalized for each participant as the middle-ranked (fourth) juice during a pre-conditioning rating session of seven different juices. Participants rated this juice, as well as their third- and fifth-ranked juices after five and ten conditioning sessions, using a 100-mm line scale anchored with extremely unpleasant to extremely pleasant. For the purposes of data analysis, participants were categorized as post-lunch caffeine users or not, based on food diaries completed prior to the

conditioning trials. Rated pleasantness of the target juice increased for the users when it had been paired with caffeine compared to when it had been paired with placebo. In contrast, there was no change in rated pleasantness for the target juice for the nonusers, regardless of whether it had been paired with caffeine or placebo. Liking for the control juices (previously ranked three or five) also failed to change for both types of users, indicating that the effect was specific to the caffeine-paired juice for the users.

It could be argued that users in this experiment were in a relatively deprived state, because they did not have their customary post-lunch caffeinated beverage; however, caffeine deprivation state was not explicitly manipulated. Yeomans and colleagues (e.g., Chambers, Mobini, & Yeomans, 2007; Yeomans, et al., 2000; Yeomans, Pryke, & Durlach, 2002; Yeomans, Spetch, & Rogers, 1998) developed a within-day two-stage procedure in order to explicitly manipulate caffeine deprivation state and investigate its effect on caffeine conditioning of flavor liking. In these experiments, participants were asked to refrain from eating or drinking anything but water from 2300 the night before each session and to arrive at the lab in the morning for breakfast (cereal and tea). The tea provided at breakfast was used to manipulate the deprivation state of participants for the second stage of the experiment each day, which occurred 120 minutes later. In other words, the tea provided at breakfast either did or did not contain caffeine, depending on the caffeine deprivation state required for the condition to which the participant was assigned. During the subsequent second stage, participants consumed a novel-flavored juice, which similarly contained caffeine or not, depending on the condition to which the participant was assigned. Sessions included pre- and post-consumption mood, thirst, and hunger ratings, and ratings of the beverages for aspects such as novelty, pleasantness, sweetness, and sourness. Participants were also required to give saliva samples to encourage compliance with abstention requirements, although these were not actually analyzed, because individual differences in caffeine metabolism make it impossible to establish abstention after only nine to ten hours. Prior to recruitment for any specific experiment, participants completed multiple questionnaires about their eating habits and preferences, which included questions about caffeine usage along with questions about other food ingredients or additives. Data from these questionnaires were used to select participants according to their estimated average caffeine consumption. Target participants were subsequently sent a specific recruitment letter inviting them to participate in a study on the effects of commonly consumed

TABLE 1. Design of experiment by Yeomans, Specht, and Rogers (1998).

Condition	*Breakfast*	*Mid-morning*
CC	Herbal Tea with Caffeine	Juice with Caffeine
CN	Herbal Tea with Caffeine	Juice without Caffeine
NC	Herbal Tea without Caffeine	Juice with Caffeine
NN	Herbal Tea without Caffeine	Juice without Caffeine

beverages on mood, and included a list of five potential drink ingredients, including caffeine, which they might be asked to consume. If they had contraindications for any of these ingredients, or met particular health exclusions, they were advised not to take part. The intention was to mask caffeine and liking as the focus of the study. Post-study debriefings were given to determine the extent to which this masking was successful.

This two-stage procedure was first used to establish whether being caffeine deprived was necessary to establish increased flavor ratings for flavors paired with caffeine in habitual caffeine consumers (at least 195 mg/day). Table 1 illustrates the design of this experiment (Yeomans, Specht, & Rogers, 1998). Four groups of participants were distinguished by whether they received caffeine (100 mg) with their tea at breakfast (CC and CN) or not (NC and NN), and whether they received caffeine with their juice mid-morning (CC and NC) or not (CN and NN). From day one to day four of the experiment, rated pleasantness of the tea served at breakfast increased in groups CC and CN, but not in groups NC and NN, demonstrating the basic flavor conditioning effect in deprived caffeine users. The same pattern was seen for the juice, comparing groups NC and NN. The novel condition is CC, which did receive caffeine in the juice, so was less deprived than NC, having received caffeine in the tea at breakfast. Neither CC nor CN, which also received caffeine in the tea at breakfast, showed any change in rated pleasantness of the juice, whether the juice was paired with caffeine (CC) or not (CN).

These results showed flavor liking increases when habitual caffeine consumers receive flavor-caffeine pairings while caffeine deprived, but not while caffeine replete. This led us to wonder if the expression of the acquired enhanced liking also depends on caffeine deprivation state. Just as the aroma of food might seem more attractive when hungry than when sated, a flavor associated with the effects of caffeine might seem more attractive when caf-

feine deprived than when caffeine replete. To test this possibility, Yeomans et al. (2000) examined whether participants who had consumed a caffeine-paired flavor while deprived and expressed an enhanced liking for that flavor during conditioning, would continue to express that enhanced liking when no longer deprived. The experiment used the two-stage design described above. During the conditioning phase the breakfast tea was never paired with caffeine and two groups differed as to whether their juice contained caffeine (C) or placebo (P). Compared with the first conditioning trial, as expected, pleasantness ratings for the juice on the fourth conditioning trial were higher in C, but were lower in P. On the fifth session, half of the participants in each condition received caffeine in their breakfast tea for the first time. Thus, on the fifth session, half the participants rated the juice while still caffeine deprived as on prior trials (Deprived-C and Deprived-P), but the rest rated the juice after a caffeine pre-load in the tea (Preload-C and Preload-P). Not surprisingly, participants' ratings of the juice on day five in Deprived-C and Deprived-P were similar to their ratings on day four. In contrast, in the pre-load conditions, Preload-C expressed a significant decrease in juice rating compared to day four, and Preload-P expressed a significant increase in juice rating compared to day four.

The first new finding from this study was that the increase in rated liking for the caffeine-paired flavor was not expressed when participants were no longer caffeine deprived.This finding is in accordance with others that demonstrate that hedonic responses are modulated by physiological state. For example, rated pleasantness of food is higher when hungry than when sated (Booth, Mather, & Fuller, 1982; Cabanac, 1971; Johnson & Vickers, 1993), and an acquired preference for a flavor paired with a high-protein lunch is abolished when the protein-paired flavor is evaluated after a high-protein preload (Gibson, Wainwright, & Booth, 1995).

The second new finding from the Yeomans, et al. (2000) experiment was that the decrease in rated liking for a flavor repeatedly experienced in a state of caffeine deprivation (Preload-P) was not expressed when participants were no longer caffeine deprived. It is possible that this reversal of the acquired dislike for the flavor repeatedly experienced while caffeine deprived was nonspecific. Participants in the Preload-P condition might have been in an enhanced mood (as a result of the preload), compared with their prior experiences at the lab. This contrast in mood, compared to what they had previously experienced could have resulted in a nonspecific increase in rated liking of

the juice. However, Yeomans, Pryke, & Durlach (2002) tested this possibility and found specificity. That experiment replicated the findings that pleasantness ratings for a caffeine-free drink experienced in a caffeine-deprived state decreased; and that this decrease was reversed when the drink was rated after a caffeine preload. But the ratings of other drinks were not elevated in the preload condition compared with the still deprived condition. The pattern of results indicates that both the learned enhancement and learned depression of liking for a flavor based on its relationship to caffeine are expressed only in a state of caffeine deprivation.

We also examined whether people learn a flavor-caffeine association when not caffeine-deprived. We had already established that under these circumstances, no increase in rated pleasantness occurs; but the question was, what would happen if people undergoing this experience were subsequently asked to rate the flavor when they were caffeine deprived. If an association learned between a flavor and caffeine while caffeine deprived is not expressed when caffeine replete, perhaps the converse would also occur. It is clear that animals can learn about events that are irrelevant to their current motivational state. For example, rats that consume a flavored salty drink learn an association between the flavor and the taste of salt; but, to demonstrate this in behavior it is necessary to endow the salt with significance. Creating an aversion to the salty taste by pairing it with a LiCl injection makes the flavor aversive as well (Rescorla & Durlach, 2011). Conversely, chemically inducing a salt-need makes a salt-paired stimulus attractive (e.g., Dayan & Berridge, 2014).

To investigate this question, Yeomans, et al. (2001) again used the two-stage design. The latent learning group (LL) was given caffeine both in their breakfast tea and their mid-morning juice each day, except on the critical final test day. On that day, the breakfast tea did not contain caffeine, so that they rated the juice mid-morning in a relatively deprived state. The latent learning control group (LL-Control) received the identical treatment, except they were given caffeine in their breakfast tea, just like during the previous sessions (same as CC in Table 1). Two other control groups were included to assess the extent of any latent learning revealed on the test day. One group was the standard conditioning condition, equivalent to NC in Table 1. The other group was the standard conditioning control, equivalent to NN in Table 1. As expected, ratings of the juice increased over trials in the NC condition, and decreased in the NN condition. No changes in rated pleasantness of the juice were observed in the LL or LL-Control groups. This remained

the case even on the final test day when LL was tested in a deprived state. Thus switching caffeine deprivation state in the LL condition failed to reveal latent learning of a caffeine association.

We also investigated the caffeine conditioning effect in self-selected habitual caffeine consumers who were chronically withdrawn from caffeine (Tinley, Yeomans, & Durlach, 2003). The data presented thus far suggest that flavor conditioning by caffeine only occurs under conditions in which the caffeine reverses short-term withdrawal effects. It does not occur in self-selected non- or low-users, nor in regular users when overnight withdrawal has been compensated for by a caffeine preload. Nevertheless, it is possible that in the absence of a caffeine preload or acute withdrawal, caffeine might have some positive reinforcing effects; but these might be difficult to detect after habitual use. Self-selected users might become tolerant to caffeine's positive effects, such that it's reinforcing power is only evident during acute withdrawal. On the other hand, chronic withdrawal might remove tolerance and allow a positive reinforcing effect of caffeine to be detected.

If caffeine can have reinforcing effects upon initial use, then individual differences in caffeine's effects in relatively naïve users could provide an explanation for subsequent caffeine consumption habits. If this were the case, people with an initial positive response should be more likely to adopt it, and those with an initial negative response should be less likely to adopt it. There is evidence that a sub-population finds the effects of caffeine unpleasant and that this may have a genetic basis (Childs, et al., 2008; Goldstein, Keiser, & Whitby, 1969; Kendler & Prescott, 1999). However, Rogers et al. (2010) failed to find a relation between a polymorphism associated with an anxiogenic response to caffeine and actual caffeine consumption — evidence which does not favor a role for the initial reinforcing effects of caffeine as influential in caffeine consumption.

To address this question, Tinely, Yeomans, & Durlach (2003) tested caffeine users who were either chronically withdrawn (two weeks) or not. This was accomplished by asking participants to replace their normal coffee or tea with supplies provided, either caffeinated (maintained) or decaffeinated (withdrawn), and to refrain from consumption of any other caffeinated products. After two weeks, withdrawn participants should have overcome any withdrawal symptoms (Juliano & Giffiths, 2004), and have no trace of caffeine in their systems. Analysis of saliva was used to eliminate noncompliant participants in the withdrawn (W) condition. For conditioning days, partici-

TABLE 2

Conditioning	*Same*	*Opposite*
Deprived-Paired	Increase	No change
Deprived-Unpaired	Decrease (or No change)	No change
Preload-Paired	No change	No change
Preload-Unpaired	No change	(no data)

Note. Effect of rated liking for a novel-flavored beverage compared to its initial pre-conditioning rating in habitual caffeine users, depending on whether the beverage was repeatedly paired with caffeine during conditioning (Paired) or not (Unpaired), whether the users were overnight caffeine-deprived (Deprived) or not (Preload) during conditioning, and whether the rating was made in the Same or the Opposite state as during the conditioning.

pants were instructed to refrain from consuming anything but water from 23:00 the night before and to visit the lab between 08:30 and 09:30. Therefore participants in the maintained condition (M) were overnight withdrawn. At the lab, participants were given breakfast with a novel herbal tea. Half of the participants received 70 mg caffeine in the tea (M-Caff and W-Caff), whereas the rest did not (M-Placebo and W-Placebo). Participants in the M-Caff condition showed the expected increase in liking for the herbal tea over conditioning trials compared to the M-Placebo condition. However, the reverse was the case in the chronically withdrawn condition. Participants in the W-Caff condition rated the herbal tea significantly less pleasant than participants in the W-Placebo condition. Thus this experiment provided no evidence for a positive reinforcing effect of caffeine in regular caffeine users who had been chronically caffeine-abstinent. It seems likely therefore that initial adoption of caffeine containing products is not based on the reinforcing effects of caffeine, but rather on flavor-flavor, flavor-nutrient, or socio-cultural reinforcement (Sheperd & Raats, 2006; Zellner, 1991).

Table 2 summarizes some of the results presented thus far. In regular caffeine consumers who are overnight withdrawn, liking for a novel flavor paired with caffeine increases; but this increase is not evident if tested after a caffeine preload (row one). In regular caffeine consumers who are overnight withdrawn, liking for a novel flavor not paired with caffeine either decreases or fails to change. If it does decrease, that distaste disappears if subsequently assessed after a caffeine preload (row two). In regular caffeine consumers who

are overnight withdrawn, but receive a caffeine preload, liking for a novel flavor paired with caffeine fails to change, and is not evident even if subsequently assessed without the preload (row three); i.e., no latent learning is apparent.

One interpretation of these results is that they are an example of Pavlovian flavor-consequence learning (Yeomans, 2006). That is, the learned liking might be based on consumers learning that consuming the flavor is followed by positive affective consequences. An acquired liking is supported by the association of the flavor with withdrawal relief. When withdrawal relief is not required, the liking is not expressed. This would be analogous to a rat displaying disinterest in a conditional stimulus (CS) previously paired with food when no longer hungry. Latent learning does not occur, because withdrawal relief is not experienced in association with the flavor in the latent learning condition. The flavor-consequence interpretation seems less plausible for the state dependency of dislike resulting from association of a flavor with acute withdrawal, however (Table 2, row two). An association between the flavor and the negative affect of acute withdrawal could account for the learned distaste; but it is not clear that this should be reversed when liking for the flavor is measured after a caffeine preload. When not in a state of acute withdrawal, cues associated with that state should still be aversive.

Another possible interpretation of these results is that they are an example of incentive salience learning, as described by Anselme and Robinson (2015), Berridge (2012), and Dayan and Berridge (2014). Incentive salience has been equated with the psychological state of wanting or craving (Anselme & Robinson, 2015; Berridge, 2012; Dayan & Berridge, 2014). The claim is that while wanting and liking often co-occur, they depend on different brain regions and can be dissociated by neurobiological and/or behavioral manipulations. It has been suggested that the feelings and underlying mechanisms associated with craving rather than liking are responsible in large part to drug addiction and other compulsive behaviors. With respect to caffeine, craving in and of itself, has received little scientific investigation (Juliano and Griffiths, 2004).

According to Berridge and colleagues, conditioned craving is automatically and immediately affected by a change in motivational state. The sensitivity of both the learned liking and the learned disliking as represented in rows one and two of Table 2 therefore fit with the possibility that in the caffeine experiments, participants' pleasantness ratings were a reflection of this type of learning.The case for latent learning is somewhat less clear. Dayan and Berridge (2014) conceive of changes in physiological and neurobiologi-

cal factors as impacting the stimulus representation of the unconditional stimulus (UCS), and thereby also the value of any associated CS. If the caffeine given to participants in the mid-morning juice in the latent learning condition had any stimulus effects, then latent learning would be predicted to occur. On the other hand, if the caffeinated juice consumed after having had caffeine at breakfast had no appreciable effects, there would be no UCS representation. It would be as if there were no caffeine in the mid-morning juice, and participants in the LL condition should respond the same when tested caffeine deprived as participants treated like condition CN in Table 1, were they subsequently tested while deprived. Unfortunately, this condition has not been examined (row 4, Table 2). Yet there is no reason to suspect that such a condition would be other than neutral to the juice's flavor, having never had it paired with caffeine or with caffeine withdrawal.

The failure to find latent learning could be interpreted as additional support for the claim that caffeine has no state-independent reinforcement power, but rather that the reinforcing effects of caffeine can be reduced to its ability to alleviate symptoms of withdrawal. This issue might only be settled if caffeine-naïve participants can be tested. In rats, caffeine has been shown to support flavor preferences or aversions, depending on the caffeine dose paired with the flavor (Fedorchat, Mesita, & Plater, 2002; Myers & Izbicki, 2006); however, in all cases in which a preference has been demonstrated, at the time of testing the rats had not consumed caffeine for one to two days and so were likely in a state of caffeine withdrawal. Animal experiments in which a caffeine preload is given or not prior to test in order to assess the effect of caffeine deprivation level on preference have not been conducted.

The research discussed in this chapter has a bearing on the role of contingency awareness in human Pavlovian conditioning (e.g., Hütter et al., 2012; Lovibond & Shanks, 2002; Ruys & Stapel, 2009; Schultz & Helmstetter, 2010). A controversy exists as to whether conditioning can occur if the participant is unaware of the stimulus contingencies. For evaluative conditioning, explicit knowledge of the CS-UCS relationship has been shown to be important in observing the effect (e.g., Hofman, et al., 2010); however, debate still exists based on methods of establishing participant awareness or lack of awareness and other stimulus factors (Hütter et al., 2012; Ruys & Stapel, 2009). In the present experiments, except for the one study on long-term caffeine withdrawal, the experimenters made an effort to mask the purpose of the studies by telling participants the research was on mood (as opposed to liking). Post-

study debriefings were given to determine the extent to which this masking was successful. The results for the three two-stage experiments discussed in detail here are presented in Table 3. During these debriefings participants were first asked an open-ended question about the purpose of the study. For this initial question only ten of the 132 participants tested (less than 8%) spontaneously mentioned caffeine; and only one out the 132 mentioned flavor rating. All others mentioned mood. Following this initial probing participants were asked explicitly whether they thought the study involved caffeine. A higher percentage responded yes to this, 32%. Finally, participants were told that the study did involve caffeine and asked them whether they thought they received caffeine or if they could identify which drinks, if any, contained caffeine. Across the three experiments 79 participants received caffeine in their mid-morning drink, the drink of interest in terms of rated liking. Among them, 28, or about 35%, correctly identified that drink as containing caffeine. However, across the three experiments there were 53 participants who did not receive caffeine in their mid-morning drink. Among them, 19, or about 36%, incorrectly identified that drink as containing caffeine. Thus, whether the target drink contained caffeine or not, the percent of participants saying that it

Table 3. Participant responses to debriefing questions across experiments. See text for explanation of conditions

	Yoemans, Specht & Rogers (1998)			
Condition	*CC (N=9)*	*CN (N=9)*	*NC (N=9)*	*NN (N=9)*
Spontaneously mentioned caffeine	1	2	0	1
Spontaneously mentioned mood	9	9	9	9
Spontaneously mentioned flavor perception or liking	0	0	0	0
Said caffeine was involved when explicitly asked	5	5	3	3
Correctly identified target drink as having caffeine or not	2	–	4	–
Incorrectly identified target drink as having caffeine	–	2	–	1

(Continued)

Condition	*Yeomans, et al. (2000)*			
	Deprived-P (N=11)	*Deprived-C (N=11)*	*Preload-P (N=11)*	*Preload-C (N=11)*
Spontaneously mentioned caffeine	0	1	1	2
Spontaneously mentioned mood	1 correctly identified the purpose as having to do with rating the drinks; but group membership not reported. All others mentioned mood.			
Spontaneously mentioned flavor perception or liking				
Said caffeine was involved when explicitly asked	3	0	3	4
Correctly identified target drink as having caffeine or not	–	5	–	5
Incorrectly identified target drink as having caffeine	7	–	8	–

Condition	*Yeomans, et al. (2001)*			
	NN (N=13)	*NC (N=13)*	*LL-Control (N=13)*	*LL (N=13)*
Spontaneously mentioned caffeine	0	0	2	1
Spontaneously mentioned mood	13	13	13	13
Spontaneously mentioned flavor perception or liking	0	0	0	0
Said caffeine was involved when explicitly asked	3	4	5	5
Correctly identified target drink as having caffeine or not	–	5	3	4
Incorrectly identified target drink as having caffeine	1	–	–	–

did was about the same. Although there are known difficulties with trying to establish contingency awareness via retrospective self-report, these caffeine conditioning data seem to represent an example of evaluative conditioning without contingency awareness.

References

Addicott, M. A., & Laurienti, P. J. (2009). A comparison of the effects of caffeine following abstinence and normal caffeine use. *Psychopharmacology*, *207*(3), 423-431. doi:10.1007/s00213-009-1668-3.

Anselme, P., & Robinson, M. F. (2015). 'Wanting,' 'liking,' and their relation to consciousness. *Journal Of Experimental Psychology: Animal Learning And Cognition*, doi:10.1037/xan0000090.

Berridge, K. C. (2012). From prediction error to incentive salience: Mesolimbic computation of reward motivation. *European Journal Of Neuroscience*, *35*(7), 1124-1143. doi:10.1111/j.1460-9568.2012.07990.x.

Booth, D. A., Mather, P., & Fuller, J. (1982). Starch content of ordinary foods associatively conditions human appetite and satiation, indexed by intake and eating pleasantness of starch-paired flavours. *Appetite*, *3*(2), 163-184. doi:10.1016/S0195-6663(82)80009-3.

Cabanac, M. (1971). Physiological role of pleasure. *Science*, *173*, 1103-1107. doi: 10.1126/science.173.4002.1103.

Chambers, L., Mobini, S., & Yeomans, M. R. (2007). Caffeine deprivation state modulates expression of acquired liking for caffeine-paired flavours. *Quarterly Journal Of Experimental Psychology*, *60*(10), 1356-1366. doi:10.1080/17470210601154545.

Childs, E., Hohoff, C., Deckert, J., Xu, K., Badner, J., & De Wit, H. (2008). Association between ADORA2A and DRD2 polymorphisms and caffeine-induced anxiety. *Neuropsychopharmacology*, *33*(12), 2791-2800. doi:0.1038/npp.2008.17.

Dack, C., & Reed, P. (2009). Caffeine reinforces flavor preference and behavior in moderate users but not in low caffeine users. *Learning and Motivation*, *40*(1), 35-45. doi:10.1016/j.lmot.2008.05.002.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience*, *14*(2), 473-492. doi:10.3758/s13415-014-0277-8.

Fedorchak, P. M., Mesita, J., Plater, S. A., & Brougham, K. (2002). Caffeine-reinforced conditioned flavor preferences in rats. *Behavioral Neuroscience*, *116*(2), 334-346. doi:10.1037/0735-7044.116.2.334.

Gibson, E. L., Wainwright, C. J., & Booth, D. A. (1995). Disguised protein in lunch after low-protein breakfast conditions food-flavor preferences dependent on recent lack of protein intake. *Physiology & Behavior*, *58*(2), 363-371. doi:10.1016/0031-9384(95)00068-T.

Goldstein, A., Kaiser, S., & Whitby, O. (1969). Psychotropic effects of caffeine in man. IV Quantitaive and qualitative differences associated with habituation to caffeine. *Clinical Pharmacology and Therapeutics*, *10*, 489-497.

Heatherley, S. V., Hancock, K. F., & Rogers, P. J. (2006). Psychostimulant and other effects of caffeine in 9- to 11-year-old children. *Journal Of Child Psychology And Psychiatry*, *47*(2), 135-142. doi:10.1111/j.1469-7610.2005.01457.x.

Heckman, M. A., Weil, E., & Gonzalez de Mejia, E. (2010). Caffeine (1, 3, 7-trimethylxanthine) in foods: A comprehensive review on consumption, functionality, safety, and regulatory matters. *Journal of Food Science*, *75*(3), 77-87. doi:10.1111/j.1750-3841.2010.01561.x.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390-421. doi:10.1037/a001891.

Hütter, M., Sweldens, S., Stahl, C., Unkelbach, C., & Klauer, K. C. (2012). Dissociating contingency awareness and conditioned attitudes: Evidence of contingency-unaware evaluative conditioning. *Journal Of Experimental Psychology: General*, *141*(3), 539-557. doi:10.1037/a0026477.

James, J. E. (1991). *Caffeine and health*. London: Academic Press.

Johnson, J., & Vickers, Z. (1993). Effects of flavor and macronutrient composition of food servings on liking, hunger, and subsequent intake. *Appetite*, *21*(1), 25-39. doi:10.1006/appe.1993.1034.

Juliano, L. M., & Griffiths, R. R. (2004). A critical review of caffeine withdrawal: Empirical validation of symptoms and signs, incidence, severity, and associated features. *Psychopharmacology*, *176*(1), 1-29. doi:10.1007/s00213-004-2000-x.

Kendler, K. S., & Prescott, C. A. (1999). Caffeine intake, tolerance, and withdrawal in women: A population-based twin study. *American Journal of Psychiatry*, *156*(2), 223-228.

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3-26. doi:10.1037/0097-7403.28.1.3.

Mitchell, D. C., Knight, C. A., Hockenberry, J., Teplansky, R., & Hartman, T. J. (2014). Beverage caffeine intakes in the U.S. *Food and Chemical Toxicology*, *63*, 136-142. doi:10.1016/j.fct.2013.10.042.

Myers, K. P., & Izbicki, E. V. (2006). Reinforcing and aversive effects of caffeine measured by flavor preference conditioning in caffeine-naive and caffeine-acclimated rats. *Physiology & Behavior*, *88*(4-5), 585-596. doi:10.1016/j.physbeh.2006.05.010.

Rescorla, R. A., & Durlach, P. J. (1981). Within-event learning. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 81-111). Hillsdale, NJ: Erlbaum.

Richardson, N. J., Rogers, P. J., & Elliman, N. A. (1996). Conditioned flavour preferences reinforced by caffeine consumed after lunch. *Physiology & Behavior*, *60*(1), 25-263. doi:10.1016/0031-9384(95)02203-1.

Rogers, P. J. (2000). Why we drink caffeine-containing beverages, and the equivocal benefits of regular caffeine intake for mood and cognitive performance. *Caffeinated Beverages, ACS Symposium Series, Volume 754*, 37-45. doi:10.1021/bk-2000-0754.ch005.

Rogers, P. J. (2014). Caffeine and alertness: In defense of withdrawal reversal. *Journal of Caffeine Research*, *4*(1), 3-8. doi:10.1089/jcr.2014.0009.

Rogers, P. J., Hohoff, C., Heatherley, S. V., Mullings, E. L., Maxfield, P. J., Evrshed, R. P., Deckert, J., & Nutt, D. J. (2010). Association of the anxiogenic and alerting effects of caffeine with ADORA2A and ADORA1 polymorphisms and habitual level of caffeine consumption. *Neuropsychopharmacology*, *35*(9), 1973-1983. doi: 10.1038/npp.2010.71.

Ruys, K. I., & Stapel, D. A. (2009). Learning to like or dislike by association: No need for contingency awareness. *Journal of Experimental Social Psychology*, *45*(6), 1277-1280. doi:10.1016/j.jesp.2009.06.012.

Shepherd, R., & Raats, M. (2006). *The Psychology of Food Choice*. Cambridge, MA: CABI. doi:10.1079/9780851990323.0000.

Schuh, K. J., & Griffiths, R. R. (1997). Caffeine reinforcement: The role of withdrawal. *Psychopharmacology*, *130*(4), 320-326. doi:10.1007/s002130050246.

Schultz, D. H., & Helmstetter, F. J. (2010). Classical conditioning of autonomic fear responses is independent of contingency awareness. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(4), 495-500. doi:10.1037/a0020263.

Tinley, E. M., Durlach, P. J., & Yeomans, M. R. (2004). How habitual caffeine consumption and dose influence flavour preference conditioning with caffeine. *Physiology & Behavior*, *82*(2-3), 317-324. doi:10.1016/j.physbeh.2004.03.018.

Tinley, E. M., Yeomans, M. R., & Durlach, P. J. (2003). Caffeine reinforces flavour preference in caffeine-dependent, but not long-term withdrawn, caffeine consumers. *Psychopharmacology*, *166*(4), 416-423. doi:10.1007/s00213-002-1354-1.

Yeomans, M. R. (2006). The role of learning in the development of food preferences. In R. Shepherd, & M. Raats (2006). *The psychology of food choice* (pp. 93-112). Cambridge, MA: CABI. doi:10.1079/9780851990323.0000.

Yeomans, M. R., Jackson, A., Lee, M. D., Nesic, J., & Durlach, P. J. (2000). Expression of flavour preferences conditioned by caffeine is dependent on caffeine deprivation state. *Psychopharmacology*, *150*(2), 208-215. doi:10.1007/s002130000405.

Yeomans, M. R., Pryke, R., & Durlach, P. J. (2002). Effect of caffeine-deprivation on liking for a non-caffeinated drink. *Appetite*, *39*(1), 35-42. doi:10.1006/appe.2001.0480.

Yeomans, M. R., Spetch, H., & Rogers, P. J. (1998). Conditioned flavour preference negatively reinforced by caffeine in human volunteers. *Psychopharmacology*, *137*(4), 401-409. doi:10.1007/s002130050636.

Zellner, D. A. (1991). How foods get to be liked: Some general mechanisms and some special cases. In R. C. Bolles (Ed.), *The hedonics of taste* (pp. 199-217). Erlbaum: Hillsdale, N. J.

Successive Contrast Effects in a Navigation Task with Rats

Apolonia Manchón, Marta N. Torres,
Teresa Rodrigo, V. D. Chamizo*
Universitat de Barcelona, Spain

Abstract. In a set of three experiments rats were required to escape from a circular pool by swimming to a hidden platform that was located in the same place relative to a single object (i.e., a beacon). Experiment 1 trained female rats with two transparent platforms, one at water level and the other one 3 cm below water level, and established that they acted on their performance as two rewards of different magnitude (large and small, respectively). In Experiment 2 experimental female rats were exposed either to a shift from finding the platform at water level to finding it 3 cm below water level (Experiment 2a – negative contrast) or to a shift from finding the platform 3 cm below water level to finding it at water level (Experiment 2b – positive contrast), while the control groups did not have a platform shift (it was always at water level or 3 cm below water level). A clear negative contrast effect was found. Finally Experiment 3, with four groups of male rats, directly compared the two conditions and found a negative contrast effect only. These results will allow us to address an associative analysis of true spatial learning (O'Keefe & Nadel, 1978) to basic instrumental phenomena in future work, as Mackintosh would have suggested.

Introduction

Working with rats in a straight runway, Crespi (1942) is often considered as the author to provide the first parametric investigation addressing the effects of a shift in the quantity of reward (for reviews see Mackintosh, 1974; Flaherty, 1966, 1982). In Crespi's experiments (see also Zeaman, 1949) the animals were

* This research was supported by grants from the Spanish Ministry of Science and Innovation (Refs. psi2010-20424 and psi2013-47430-p) to V.D.C.

initially trained with one magnitude of reward and then shifted to another. He found that these changes in magnitude of reward were followed by important shifts in performance. For example, a reduction in the magnitude of reward produced a surprising decrease in running speed: the rats shifted from a large to a small amount of food run less vigorously to the small reward than animals that had never received the large reward (a finding that is nowadays known as *successive negative contrast effect*). If the successive change in magnitude (or quality or delay) of reward is in the opposite direction (like for example shifting the rats from a small to a large amount of food), the reverse could be expected: more vigorous running to the large reward in the shifted rats than animals that have never received the small reward (a finding that is nowadays known as *successive positive contrast effect*). This second effect, positive contrast, is harder to find and much less frequently reported in the literature (although see Maxwell, Calef, Murray, Shepart, & Norville, 1976; Mellgren, 1971, 1972; Shanab & Ferrell, 1970; Shanab, Sanders, & Premak, 1969).

In the study by Mellgren (1972), also with rats and a straight runway, animals were divided into four groups. Two of them received a small reward (2 pellets of food) when they reached the end of the runway, while the other two groups received a larger reward (22 pellets of food). A special "parameter" in this study was that the presentation of the food was always delayed 20 sec so that the rats did not run at their maximum speed. After 11 days of training in phase 1, one group of rats with each reward quantity was shifted to the alternate quantity (i.e., from 2 to 22, and from 22 to 2). The remaining two groups continued to receive the same amount of reward in this new phase, phase 2 (i.e., always 2 or 22 pellets of food). The main results found were that following a shift in magnitude of reinforcement, important shifts in performance were obtained. Rats shifted from a small to a large reward run faster for the large reward than rats that always received the large reward. In addition, rats shifted from a large to a small reward run more slowly for the small reward than rats that always received the small reward. Therefore, both successive positive and negative contrast effects were found. Then Experiment 2 of this study (Mellgren, 1972), with only an increase in reward magnitude (1-8 pellets) and an unshifted control group (the two groups receiving delayed reward), confirmed the positive contrast effect observed in Experiment 1. Mellgren discussed the negative contrast effect in terms of frustration, while the positive contrast effect was explained in terms of the possible inhibition generated by the use of delayed reinforcement.

Successive contrast effects have been found not only in appetitive instrumental conditioning but also, although less frequently, when dealing with aversive behaviour (Cándido, Maldonado, Megías, & Catena, 1992; McAllister, McAllister, Brooks, & Goldman, 1972). In the study by McAllister et al. (1972), the amount of reward was measured in terms of the contextual difference between two compartments: a danger compartment (where a warning signal was followed by shock) and a safe compartment (where neither the warning signal nor the shock were present). When the two compartments were clearly different (large reward), rats escaped from the danger compartment much more quickly than when they were similar (small reward). A change in condition from large reward to small reward led to a successive negative contrast effect. The study by Cándido et al. (1992), also with rats and one way signalled avoidance learning, revealed that time spent in a safe compartment may act similarly to magnitude, quality, or delay of reward in appetitive paradigms. Specifically, a reliable impairment of the avoidance response was obtained by suddenly decreasing the time spent in the safe compartment, from 30 sec to 1 sec — i.e., negative contrast (see also Torres et al., 2005). Complementarily, a reliable improvement of the avoidance response has been obtained by suddenly increasing the time spent in the safe compartment, from 1 sec to 30 sec — i.e., positive contrast (Cándido, Maldonado, Rodríguez, & Morales, 2002).

No studies demonstrating contrast effects have been reported with rats and a navigation task when dealing with aversive behaviour (with an appetitive task, see Pecoraro, Timberlake, & Tinsley, 1999). The present series of experiments sought to extend the successive contrast effects to a different experimental paradigm, the Morris pool, in an attempt to generalize the previous results with aversive behaviour to the spatial domain. In all the present experiments we used a circular pool full of opaque water from which the animals could escape by climbing to an invisible platform, whose location was defined in terms of a single beacon. The platform, which was transparent, was placed either at water level or 3 cm below water level, so that it allowed the rats to escape more or less from the water (large reward and small reward, respectively). In other words, we initially sought to explore whether the two platform placements (platform at water level and platform 3 cm below water level) may act similarly to two different magnitudes of reward in order to answer the following question. Would successive changes in depth of the platform relative to the level of the water produce corresponding shifts in perfor-

mance, specifically in time to reach the platform? For example, would rats shifted from a platform at water level to a platform 3 cm below water level spend more time to reach the platform than animals that had never found the platform at water level? (A successive negative contrast effect). Complementarily, would rats shifted from a platform 3 cm below water level to a platform at water level spend less time to reach the platform than animals that had never found the platform 3 cm below water level? (A successive positive contrast effect). The aim of the present study was to answer these questions.

We conducted several preliminary experiments (with 40 rats in total) to ensure that they could not see the platform when at water level; to see tentatively that the rats' performance is sensitive to the reciprocal change between the two platforms, at water level and 3 cm below water level; and finally to determine the necessary amount of escape training with the two platforms in order to produce differential asymptotic levels. Rats were trained with a single beacon, which was always placed approximately 15 cm above the water level.

Experiment 1

Rats, good swimmers but not very fond of water, quickly learn to escape from the water by swimming directly to a platform from different points of the pool (Morris, 1981). An important question to answer is would the rats' performance be different to two platforms, one at water level, Pt0 (i.e., therefore allowing them to get their body completely out of the water (large reward), the second one 3 cm below water level, Pt-3 (i.e., therefore, not allowing the rats to get their body completely out of the water (small reward)? In Experiment 1, a preliminary experiment, the relative depth of the platform from the level of the water was varied in two groups of rats in the hope that these differences may act similarly to two different magnitudes of reward (i.e., at water level, a large reward and 3 cm below water level, a small reward). Consequently, two different asymptotic levels in the time to reach the platform were expected: lower latencies in a group of rats with the platform at water level than in the second group of animals with the platform 3 cm below water level (i.e., Pt0 and Pt-3, respectively). Would that be the case? The platform always maintained a constant relationship with a single beacon (i.e., a specific object placed approximately 15 cm above the water level).

Method

Subjects

The subjects were 32 naive female Long Evans rats, approximately three months old at the beginning of the experiment. They were divided into two groups (of 16 rats each): Group Pt-3 and Group Pt0. The animals were housed in standard cages, 25 × 15 × 50 cm, in groups of two and were maintained on ad lib food and water, in a colony room with a 12:12 hr light-dark cycle. They were tested within the first 8 hours of the light cycle.

Apparatus

The apparatus was a circular swimming pool made of plastic and fiberglass and modeled after that used by Morris (1981). It measured 1.58 m in diameter and 0.65 m deep, and it was filled to a depth of 0.49 m with water rendered opaque by the addition of 1 cl/l of latex. The water temperature was maintained at 22 + 1°C. The pool was situated in the middle of a large room and mounted on a wooden platform 0.43 m above the floor. The pool was surrounded by black curtains reaching from the ceiling to the base of the pool and forming a circular enclosure 2.4 m in diameter. A circular platform, 0.11 m in diameter and made of transparent Perspex, was mounted on a rod and base, and could be placed in one quadrant of the pool, 0.38 m from the side. For one group (Group Pt-3) the top of the platform was situated 3 cm below the surface of the water, for the second group (Group Pt0) the top of the platform was situated at water level. In order to ensure that the rats do not use any inadvertently remaining static room cues to find the platform (like noises from pipes and air conditioning), the platform was semi-randomly rotated with respect to the room (90°, 180°, 270°, or 360°) with the restriction that all four positions of the room were used each day. A closed-circuit video camera with a wide-angle lens was mounted 1.75 m above the centre of the pool inside the false ceiling, and its picture was relayed to recording equipment in an adjacent room. The position of the platform was defined by a beacon (i.e., an object hanging from the ceiling, placed exactly above the invisible platform, about 15 cm above the water level). This object was a white plastic cylindrical pot, 18 cm in height and 10.5 cm in diameter.

Procedure

There were two types of trials: pretraining forced trials and training trials. Pretraining forced trials consisted of placing a rat above the platform, which was submerged 3 cm below water level, during 60 sec. The beacon was not present and the rats were given two such pretraining trials in a single day. Training trials consisted of placing a rat into the circular pool with the platform and the beacon present. The rat was given 120 sec to find the platform, and once the rat had found it, it was allowed to stay on it for 30 sec. If it had not found the platform within the 120 sec, it was picked up, placed on it, and left there for 30 sec. The platform was moved from one trial to the next, and the rat was placed in the pool in a different location on each trial, as far as possible equally often on the same or opposite side of the pool from the platform, and with the platform to the right or to the left of where the rat was placed. Rats were given 24 trials during six days, at a rate of 4 trials per day. Thus, for the animals in Group Pt-3 the platform was 3 cm below water level, while for rats in Group Pt0 the platform was at water level.

Results and Discussion

Figure 1 shows the latencies to find the platform over the course of the training trials. An ANOVA conducted on these data, taking into account the variables days (1-6) and platform placement (Pt-3, Pt0) revealed that the variables days, $F(5,150) = 17.13$ ($p < 0.001$), and platform placement, $F(1,30) = 12.16$ ($p = 0.002$), were significant. No other main effect or interaction was significant (Fs < 1.0). As expected, the performance of the rats improved as days went on; and, what is crucial, the two groups reached a different asymptotic level: lower latencies were found in group Pt0 (with the platform at water level) than in group Pt-3 (with the platform 3 cm below water level). Therefore, the two platform placements (platform at water level and platform 3 cm below water level) seem to act similarly to two different magnitudes of reward (large and small, respectively).

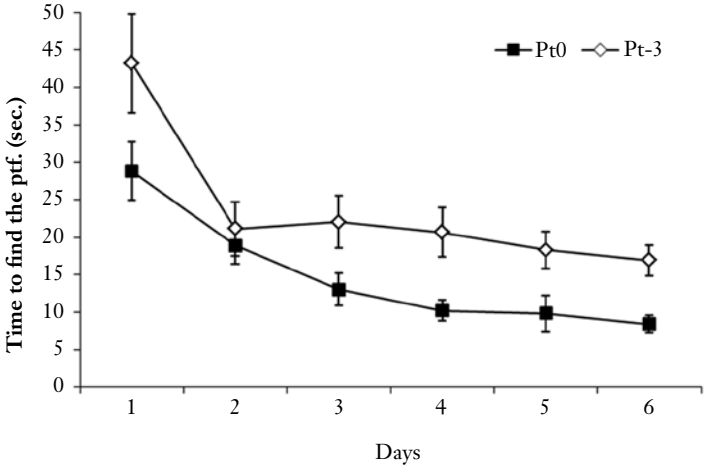

Figure 1. Mean escape latencies to find the platform for the rats of Experiment 1. The rats received four trials per day during six days. Error bars denote standard error of means.

Experiment 2

Experiment 1 established that the two platform placements (at water level and 3 cm below water level – Pto and Pt-3, respectively) could be used as two different magnitudes of reward (large and small, respectively). The two platforms produced two different asymptotic levels in the time to reach the platform, therefore allowing us to address successive contrast effects in the Morris pool. The aim of Experiment 2 was to test whether successive negative (Experiment 2a) and positive (Experiment 2b) contrast effects could be found using the two different platform placements. Both experiments (2a and 2b) had two groups of rats each (Experimental and Control) and lasted six days. For the experimental rats, the first three days (Days 1-3) were the pre-shift phase, and the rest of the days (Days 4-6), the post-shift phase. The control rats were unshifted. As in Experiment 1, the platform (either at water level or 3 cm below water level – Pto and Pt-3, respectively) always maintained a constant relationship with the single beacon.

Experiment 2a

In Experiment 2a the experimental rats underwent a change in the relative depth of the platform in the middle of the experiment (i.e., from finding the platform at water level on Days 1-3 to finding it 3 cm below water level on Days 4-6, Group Pto/-3), while the control group had the same depth of the platform during the entire experiment (i.e., the platform was always 3 cm below water level, Group Pt-3/-3). Would such a change in condition from large reward to small reward lead to a successive negative contrast effect?

Method

Subjects and apparatus
The subjects were 16 naive female Long Evans rats, approximately three months old at the beginning of the experiment. They were divided into two groups (of 8 rats each): Group Pto/-3 and Group Pt-3/-3. The animals were kept and maintained as in Experiment 1. The apparatus and the beacon, X, were the same as those used in Experiment 1.

Procedure
The general procedure was similar to that used in Experiment 1. After pretraining, the rats were given 24 trials during six days, at a rate of 4 trials per day. For Group Pto/-3 training was with the platform at water level the first three days (pre-shift phase), and with the platform 3 cm below water level the subsequent three days (post-shift phase). For Group Pt-3/-3, the platform was always 3 cm below water level.

Results and Discussion

Figure 2 (top panel) shows the latencies to find the platform over the course of the training trials. An ANOVA conducted on the pre-shift phase, taking into account the variables days (1-3) and group (Pto/-3, Pt-3/-3) revealed that, unfortunately, only the variable days was significant, $F(2,28) = 32.81$ ($p < 0.001$). No other main effect or interaction was significant (Fs < 1.0). The perfor-

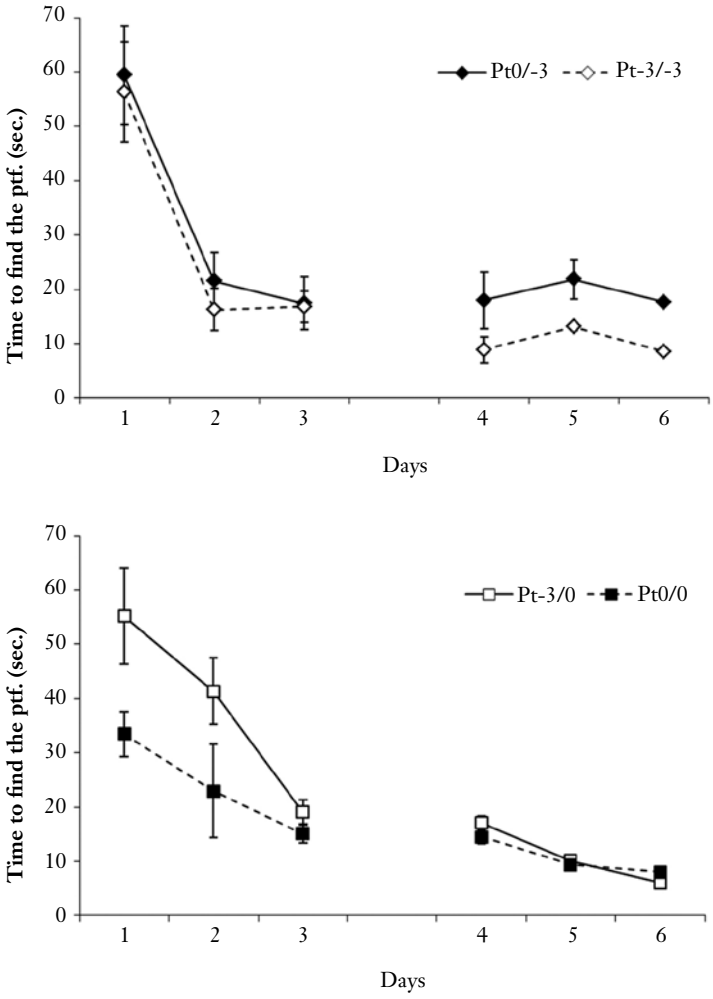

Figure 2. Top: mean escape latencies to find the platform for the rats of Experiment 2A. Bottom: mean escape latencies to find the platform for the rats of Experiment 2B. The rats received four trials per day during six days. Error bars denote standard error of means.

mance of the rats improved as days went on. An ANOVA conducted on the post-shift phase, taking into account the variables days (4-6) and group (Pto/-3, Pt-3/-3) showed that the variable group was significant, $F(1,14) = 15.61$ ($p = 0.001$), indicating a clear negative contrast effect. No other main effect or interaction was significant (Fs < 1.5). The rats shifted from a large to a small

magnitude of reward (i.e., Group Pto/-3) took longer to reach the platform than animals that had never received the large reward (i.e., Group Pt-3/-3).

Experiment 2b

As in Experiment 2a, in Experiment 2b the experimental rats underwent a change in the relative depth of the platform in the middle of the experiment (i.e., from finding the platform 3 cm below water level on days 1-3 to finding it at water level on days 4-6, Group Pt-3/0), while the control group had the same depth of the platform during the entire experiment (i.e., the platform was always at water level, Group Pto/0). Would such a change in condition from small reward to large reward lead to a successive positive contrast effect?

Method

Subjects and apparatus
The subjects were 16 naive female Long Evans rats, approximately three months old at the beginning of the experiment. They were divided into two groups (of 8 rats each): Group Pt-3/0 and Group Pto/0. The animals were kept and maintained as in the previous experiments. The apparatus and the beacon, X, were the same as those used in Experiments 1 and 2a.

Procedure
The general procedure was similar to that used in Experiment 1. After pre-training, the rats were given 24 trials during six days, at a rate of 4 trials per day. For Group Pt-3/0 training was with the platform 3 cm below water level the first three days (pre-shift phase), and with the platform at water level the subsequent three days (post-shift phase). For Group Pto/0, the platform was always at water level.

Results and Discussion

Figure 2 (bottom panel) shows the latencies to find the platform over the course of the training trials. An ANOVA conducted on pre-shift phase, taking into account the variables days (1-3) and group (Pt-3/0, Pto/0) revealed that

the variable days was significant, $F(2,28) = 12.79$ ($p < 0.001$), as well as group, $F(1,14) = 6.42$ ($p = 0.024$). No other main effect or interaction was significant (Fs < 2.0). The performance of the rats improved as days went on; and the two groups differed, indicating two magnitudes of reward, as expected. An ANOVA conducted on post-shift phase, taking into account the variables days (4-6) and group (Pt-3/0, Pt0/0) showed that only the variable days was significant, $F(2,28) = 10.72$ ($p < 0.001$). No other main effect or interaction was significant (Fs < 1.0). In conclusion, the positive contrast effect was not found.

Experiment 3

In Experiment 2, with female rats, only a negative contrast effect was found. In Experiment 3 we wondered what would have happened if the subjects had been males instead of females? The experiment consisted of four groups of males, which were conducted simultaneously. For Group Pt0/-3 training was with the platform at water level the first three days (pre-shift phase), and with the platform 3 cm below water level the subsequent three days (post-shift phase) while for Group Pt-3/-3, the platform was always 3 cm below water level (i.e., a successive negative contrast effect). For Group Pt-3/0 training was with the platform 3 cm below water level the first three days (pre-shift phase), and with the platform at water level the subsequent three days (post-shift phase) while for Group Pt0/0, the platform was always at water level (i.e., a successive positive contrast effect).

Method

Subjects and apparatus

The subjects were 32 naive male Long Evans rats approximately three months old at the beginning of the experiment. They were divided into four groups (of 8 rats each): Groups Pt0/-3 and Pt-3/0 (the experimental ones) and Groups Pt0/0 and Pt-3/-3 (the control ones). The animals were kept and maintained as in Experiments 1 and 2. The apparatus and the beacon, X, were the same as those used in the previous experiments.

Procedure

The general procedure was identical to that used in Experiment 2 (2a and 2b), although with two main exceptions. Four groups of rats (instead of two) were conducted in order to simultaneously address both a negative and a positive contrast effect. Secondly, all the animals were male rats. After pretraining, the rats were given 24 trials during six days, at a rate of 4 trials per day. For Group Pto/-3 training was with the platform at water level the first three days (pre-shift phase), and with the platform 3 cm below water level the subsequent three days (post-shift phase), while for Group Pt-3/-3, the platform was always unshifted, 3 cm below water level. Complementarily, for Group Pt-3/o training was with the platform 3 cm below water level the first three days (pre-shift phase), and with the platform at water level the subsequent three days (post-shift phase), while for Group Pto/o, the platform was always unshifted, at water level.

Results and Discussion

Figure 3 shows the latencies to find the platform over the course of the training trials. An ANOVA conducted on the pre-shift phase data, taking into account the variables days (1–3), and platform placement (platform at water level and platform 3 cm below water level), revealed that the variables days, $F(2,60) = 38.14$ ($p < 0.001$), and platform placement, $F(1,30) = 5.40$ ($p = 0.027$), were significant. No other main effect or interaction was significant ($F < 1.5$). All rats reached the platform faster as training progressed; and the rats with the platform at water level showed lower latencies (i.e., reached the platform faster) than those animals with the platform 3 cm below water level, as expected. An ANOVA conducted on the post-shift phase data, taking into account the variables days (4-6), and group (Pto/-3, Pt-3/-3, Pt-3/o, Pto/o), revealed that the variable group was significant, $F(1,28) = 5.41$ ($p = 0.005$). No other main effect or interaction was significant ($Fs < 2.0$). *A priori* simple contrasts, showed that groups Pto/-3 and Pt-3/-3 differed, $p = 0.024$, indicating a clear negative contrast effect. As in Experiment 2a, with females, the male rats shifted from a large to a small magnitude of reward (i.e., Group Pto/-3) took longer to reach the platform than animals that had never received the large reward (i.e., Group Pt-3/-3). However, groups Pt-3/o and Pto/o did not differ, $p = 0.306$, meaning that the positive contrast effect was not found.

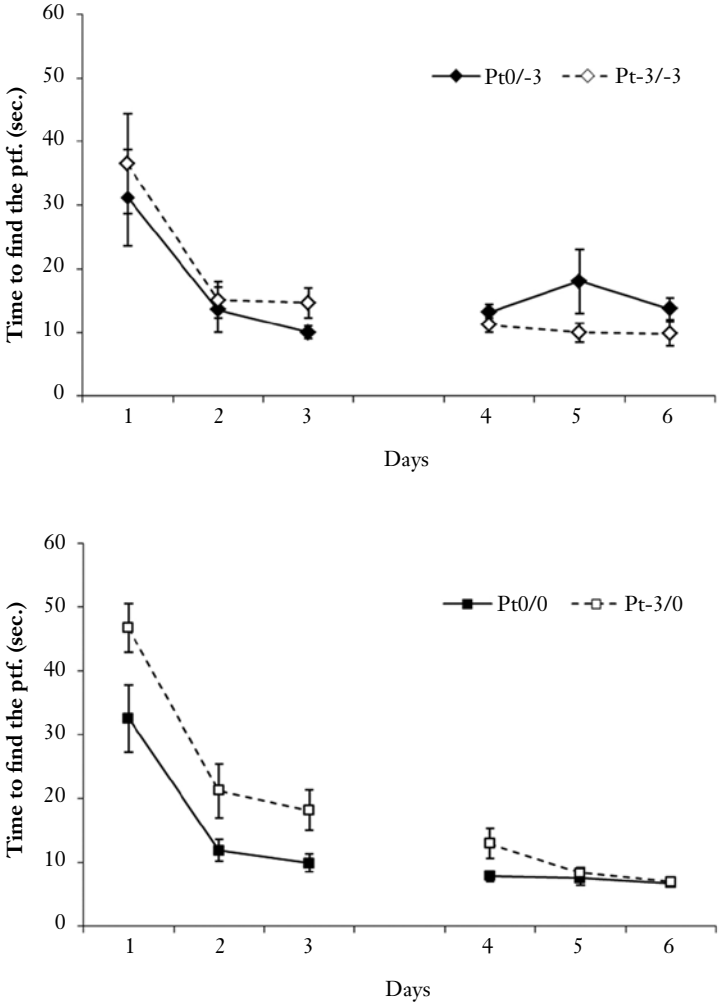

Figure 3. Top: mean escape latencies to find the platform for the rats of Experiment 3, negative contrast effect groups. Bottom: mean escape latencies to find the platform for the rats of Experiment 3, positive contrast effect groups. The rats received four trials per day during six days. Error bars denote standard error of means.

General Discussion

The results from Experiment 1 established that the two platform placements (at water level, Pt0, and 3 cm below water level, Pt-3 — i.e., which allowed

the rats to get their body completely out of the water in the first case, but not in the second case) could be used as two different magnitudes of reward (large and small, respectively), therefore allowing to undertake studying successive contrast effects in the Morris pool. Then, in Experiment 2, 2a and 2b, with two groups of female rats in each (experimental and control groups), the experimental rats underwent a change in the relative depth of the platform in the middle of the experiment, while the control rats were unshifted (i.e., Group P0/-3 and Group P-3/-3, Experiment 2a – a successive negative contrast effect; Group P-3/0, and Group P0/0, Experiment 2b – a successive positive contrast effect). The negative contrast effect was found only: After three days of training in phase 1, the rats shifted from a large to a small (i.e., from P0 to P-3) magnitude of reward reached the platform more slowly for the small reward than rats that had always received the small magnitude of reward. The successive positive contrast effect was not found: rats shifted from a small to a large (i.e., from P-3 to P0) magnitude of reward did not reach the platform faster for the large magnitude than rats that had always received the large magnitude of reward. We argued what would have happened if the subjects had been males instead of females? To answer this question was the aim of Experiment 3, which was conducted with four groups of male rats, and the same results than in females were obtained. The results showed that the negative contrast effect was significant, but that was not the case with the positive contrast effect. The performance of these animals did not differ. It could be hypothesized that this failure, both in males and in females, could be due to an artifact (i.e., a ceiling effect). This hypothesis suggests that the control rats reached the platform so fast that it was not possible for the shifted animals to reach it faster. As can be seen in the literature (Mellgren 1971, 1972; Cándido et al., 2002; Shanab et al., 1969), the successive positive contrast effect, when it happens (see Annicchiarico et al., 2016), frequently needs a special "parameter". For example, in the two experiments of the study by Mellgren (1972) this special "parameter" was that the presentation of the food was always delayed 20 sec so that the rats did not run at their maximum speed. Could we have found the positive contrast effect if the animals had been allowed to climb to the platform in a delayed manner? We do not know. Future research will have to answer this question.

The present study shows for the first time a negative contrast effect in a simple navigation task with rats. We used a circular pool full of opaque water

from which the animals could escape by climbing to a platform which was either at water level or three centimetres below the level of the water (i.e., in both cases the task implies to escape from water to motivate learning). The location of the platform was defined by a beacon. As it generally happens in appetitive tasks, we found that the larger the amount of reward (platform at water level in our case because reaching the platform allowed the animals to completely escape from the water), the faster the learning. Then, changes of the platform relative to water level (specifically, from Pt0 to Pt-3) produced a corresponding shift in performance (i.e., in time to reach the platform). We interpret the present negative contrast effects in terms of frustration (Amsel, 1992). Frustration theory has proposed that the omission or reduction of an expected appetitive reinforcer (i.e., reaching the platform in our case) is an aversive event, generating a motivational state of frustration. At present it is well known that motivational states alter the incentive value of primary reinforcers, although this is lost after extended practice (Dickinson and Balleine, 1994, 2002). In the present study, the rats of Group Pt0/-3, both males and females, that had had a large, preferred reward on Days 1-3 (so that an expectation had been made for that specific reward following their swimming behaviour on those days) experienced frustration on Days 4-6 with the less preferred reward (as evidenced by their higher swimming latencies to reach the platform on these days in comparison to the upshifted subjects, Group Pt-3/-3, both males and females). Such a frustration could not be experienced in this second group of animals because they had always found the same reward.

The contrast effects have been considered a hallmark of instrumental conditioning, where both emotional and motivational factors play a crucial role. As with appetitive tasks, in avoidance conditioning it has been shown that a longer time spent in a safe compartment improves learning (Cándido et al., 2002), while a shorter time produces an impairment of the avoidance response (Cándido et al., 1992; Torres et al., 2005). Safety signals, or the time spent in a safe place, seem to be functionally equivalent to appetitive reinforcers, acting as incentives for an avoidance response. Given the similarity between appetitive and aversive behaviour, to suggest common underlying mechanisms seems most reasonable (Dickinson, 1980; Mackintosh, 1983). The studies of contrast effects show that the effectiveness of a given reward varies with the subject's exposure to other rewards. Thus, animals adapt to a particular magnitude or quality of reward and react with increased emotion and motivation

when these magnitudes or qualities are changed. In support of this claim it has been shown that no contrast effects are observed in animals that were given tranquilizers (Maldonado, Cándido, Morales, & Torres, 2006; Morales, Torres, Megías, Cándido, & Maldonado, 1992; Torres, Morales, Cándido, & Maldonado, 1995, 1996; Torres, Morales, Megías, Cándido, & Maldonado, 1994; Rabin, 1975;), substances presumably blocking an emotional reaction.

Successive negative contrast effect has been obtained in a variety of species and preparations, as well as consummatory and Pavlovian tasks (Flaherty, 1982, 1966; Mackintosh, 1974; Papini, 2014). On the contrary the successive positive contrast effect does not seem to be equal and opposite to the negative contrast effect routinely observed when rats are shifted from a large to a small reward. The literature clearly shows that this effect is more elusive and difficult to find. In the past, Spence (1956) even suggested that only the negative case was a replicable effect. However, the study with rats by Cándido et al. (2002) offers a good relatively recent example of such an effect in one-way avoidance learning. In this study, it was found that increasing the time spent in the safe compartment (from 1 sec to 30 sec) enhanced learning of the avoidance response. The authors suggested that their results indicate that time spent in a safe context acts as a reinforcer of the avoidance response; however, its incentive value depends not only on its duration, but also on the length of the time spent in a danger compartment before the onset of the signal, as it had been hypothesized previously (Cándido, Maldonado, & Vila, 1989). Overall, Cándido et al. (2002, see also Cándido et al., 1992; Torres et al., 2005) have proposed an explanation of their results based on the modern *two-process theory of instrumental performance* (Rescorla & Solomon, 1967; Trapold & Overmier, 1972) in combination with homeostatic mechanisms. They claim that their results suggest that the avoidance response is a mixture of aversion (motivated by fear) and approach (to a safe place) behaviour. The specific weight of these components (aversion-approach) being a function of the time and the amount of activation of each emotional state (fear or relief) due to opponent homeostatic compensatory processes that occur in the danger and safe compartments during one-way avoidance learning.

In conclusion, the present results suggest that an associative analysis of spatial learning seems to apply to instrumental conditioning as well as to basic Pavlovian phenomena, as Mackintosh (1983) thought. In addition, the study adds a navigation task to the list of experimental preparations demonstrating

contrast effects. Specifically, they extend these effects to a simple escape task, when working with aversive behaviour in the spatial domain (for related work see Cándido et al., 1992; McAllister et al., 1972). A straightforward implication of the present study is that it will allow us to conduct successive contrast effects experiments, as well as other phenomena that involve surprising reward devaluation and reward omission (Papini, 2014), when dealing with multiple landmarks simultaneously, with rather more complex tasks (i.e., when learning a locale strategy instead of a taxon one, following O'Keefe & Nadel's, 1978, terminology), as Mackintosh would have suggested.

References

Amsel, A. (1992). *Frustration theory*. Cambridge, England: Cambridge University Press.

Annicchiarico, I., Glueck, A. C., Cuenya, L., Kawasaki, K., Conrad, S. E., & Papini, M. R. (2016). Complex effects of reward upshift on consummatory behavior. *Behavioural Processes*, *129*, 54-67.

Cándido, A., Maldonado, A., Rodríguez, A., & Morales, A. (2002). Successive positive contrast in one-way avoidance learning. *The Quarterly Journal of Experimental Psychology*, *55B*(2), 171-184.

Cándido, A., Maldonado, A., Megías, J. L., & Catena, A. (1992). Successive negative contrast in one-way avoidance learning in rats. *The Quarterly Journal of Experimental Psychology*, *45B*, 15-32.

Cándido, A., Maldonado, A., & Vila, J. (1989). Relative time in dangerous and safe places influences one-way avoidance learning in the rat. *The Quarterly Journal of Experimental Psychology*, *41B*, 181-199.

Crespi, L. P. (1942). Quantitative variation in incentive and performance in the white rat. *American Journal of Psychology*, *40*, 467-517.

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge: Cambridge University Press.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning and Behavior*, *22*, 1-18.

Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In R. Gallistel (Ed.), *Steven's handbook of experimental psychology. Vol. 3. Learning, motivation, and emotion* (pp. 497-533). NY: John Wiley.

Flaherty, C. F. (1966). *Incentive relativity*. Cambridge: Cambridge University Press.

Flaherty, C. F. (1982). Inventive contrast: A review of behavioural changes following shifts in reward. *Animal Learning & Behavior*, *4*, 409-440.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London: Academic Press.
Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford: Clarendon Press.
Maldonado, A., Cándido, A., Morales, A., & Torres, C. (2006). The effect of diazepam on successive positive contrast in one-way avoidance learning. *International Journal of Psychology and Psychological Therapy*, *6*(2), 249-260.
McAllister, D. E., McAllister, W. R., Brooks, C. I., & Goldman, J. A. (1972). Magnitude and shift of reward in instrumental aversive learning in rats. *Journal of Comparative and Physiological Psychology*, *80*(3), 490-501.
Maxwell, F. R., Calef R. S., Murray D. W., Shepard, F. C., & Norville, R. A. (1976). Positive and negative contrast following multiple shifts in reward magnitude under high drive and immediate reinforcement. *Animal Learning & Behavior*, *4*, 480-484.
Mellgren, R. L. (1971). Positive contrast in the rat as a function of the number of preshift trials in the runaway. *Journal of Comparative & Physiological Psychology*, *77*, 329-336.
Mellgren, R. L. (1972). Positive and negative contrast effects using delayed reinforcement. *Learning & Motivation*, *3*, 185-193.
Morales, A., Torres, C., Megías, J. L., Cándido, A., & Maldonado, A. (1992). Effect of diazepam on successive negative contrast in one-way avoidance learning. *Pharmacology Biochemistry & Behavior*, *43*, 153-157.
Morris, R. G. M. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation*, *12*, 239-260.
O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Oxford University Press.
Papini, M. (2014). Diversity of adjustments to reward downshifts in vertebrates. *International Journal of Comparative Psychology*, *27*, 420-445.
Pecoraro, N. C., Timberlake, W. D., & Tinsley, M. (1999). Incentive downshifts evoke search repertoires in rats. *Journal of Experimental Psychology. Animal behavior processes*, *25*(2), 153-167.
Rabin, J. S. (1975). Effects of varying sucrose reinforcers and anobarbital sodium on positive contrast in rats. *Animal Learning and Behaviour*, *3*, 290-294.
Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*, 151-182.
Shanab, M. E., & Ferrell, H. J. (1970). Positive contrast obtained in the Lashley maze under different drive conditions. *Psychonomic Science*, *20*, 31-32.
Shanab, M. E., Sanders, R., & Premack, D. (1969). Positive contrast in runway obtained with delay of reward. *Science*, *164*(3880), 724-725.
Spence, K. (1956). *Behavior theory and conditioning*. New Haven, CT: Yale University Press.

Torres, C., Morales, A., Megías, J. L., Cándido, A., & Maldonado, A. (1994). Flumazenil antagonizes the effect of diazepam on negative contrast in one-way avoidance learning. *Behavioural Pharmacology, 5*, 637-641.

Torres, C., Morales, A., Cándido, A., & Maldonado, A. (1995). Differential effect of buspirone and diazepam on negative contrast in one-way avoidance learning. *European Journal of Pharmacology, 280*, 277-284.

Torres, C., Morales, A., Cándido, A., & Maldonado, A. (1996). Successive negative contrast in one-way avoidance: Effect of thiopental sodium and chlorpromazine. *European Journal of Pharmacology, 314*, 269-275.

Torres, C., Cándido, A., Escarabajal, M. D., de la Torre, L., Maldonado, A., Tobeña, A., & Fernández-Teruel, A. (2005). Successive negative contrast effect in one-way avoidance learning in female roman rats. *Physiology & Behavior, 85*, 377-382.

Trapold, M. A., & Overmier, J. B. (1972). The second learning process in instrumental learning. In A. A. Black & W. E Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 427-452). New York: Appleton-Century-Crofts.

Instrumental Conditioning Revisited: Updating Dual-Process Theory

Anthony Dickinson
University of Cambridge, UK

Abstract. This chapter is an extended addendum to one on instrumental conditioning that I contributed to Mackintosh's *Animal Learning and Cognition* (1994). The central theme of that chapter was that instrumental behavior is controlled by two dissociable processes: a goal-directed and an habitual process. Here I review three aspects of the research on this dual-process theory that has emerged over the last two decades. First, behavioral control by these two processes, originally established in rodents, has been extended to human behavior. Second, goal-directed and habitual control have been doubly dissociated by neurobiological manipulations and measures. Finally, the intervening decades have seen important developments in computational and psychological theories of instrumental behavior.

Introduction

Over twenty years ago, Nicholas Mackintosh asked me to contribute a chapter to his volume *Animal Learning and Cognition* on instrumental conditioning (Dickinson, 1994), that is on behavior controlled by the contingency between an action or response and the consequent outcome or reinforcer. The central theme of the chapter was that instrumental behavior is controlled by two dissociable processes: a goal-directed and an habitual process (Dickinson, 1985). I argued that an action is goal-directed if it is mediated by the interaction of a representation of the causal relationship between the action and outcome and a representation of the current incentive value, or utility, of the outcome in a way that rationalizes the action as instrumental for attaining the goal. A role for the action-outcome representation ensures that action is *directed* at obtaining the outcome, whereas the involvement of the incentive value representation establishes that the outcome functions as a *goal*.

I offered two behavioral assays for assessing the role of each of these representations. I illustrated the first, instrumental assay by a study that assessed the sensitivity to the causal relationship between action and outcome by varying their contingency (Dickinson & Mulatero, 1989). We trained hungry rats to press two levers, one of which delivered grain pellets and the other a sugar solution at the same source, both with a probability of .05 for the first press in any second. Having established responding, we then degraded the contingency between one of the actions and its outcome by arranging that this outcome occurred with the same .05 probability in any second without a lever press. This contingency degradation had the effect of removing the causal relationship between this action and its outcome in that the rat would have received exactly the same number of these outcomes irrespective of frequency with which it pressed the associated lever. By contrast, the causal relationship remained in force for alternative lever press and its outcome in that the frequency of this outcome still depended on the number of alternative presses performed. Importantly, we observed that rats pressed less on the lever for which the contingency was degraded than on that for which contingency was maintained, thereby demonstrating that the rats were sensitive to the causal action-outcome relationship rather than just to the contiguity between action and outcome in that the probability of a paired outcome was the same for both actions.

The role of the incentive representation in instrumental performance is assessed by the reinforcer or outcome revaluation procedure, which I illustrated with a study with Adams (Adams & Dickinson, 1981). We trained rats to press a lever for one type of outcome, grain or sugar pellets, while delivering the other non-contingently. Following this instrumental training, either the contingent or non-contingent pellet was devalued by conditioning a food aversion from it in the absence of the lever, thereby reducing the incentive value of this type of pellet. If lever pressing was mediated by the incentive value of the contingent outcome, devaluing this outcome, rather than the non-contingent one, should have reduced lever pressing when the rats once again had access to the lever. This is exactly what we observed, leading us to conclude that lever pressing was goal-directed. It is important to note that the assessment of instrumental performance following outcome revaluation was conducted in the absence of the outcome, or in other words in extinction, so that responding must have been controlled by a representation of the current incentive value of the outcome rather than by the direct reinforcing or punishing effects of delivering the outcome itself.

Although this study demonstrated that simple instrumental conditioning could be goal-directed, we did not conclude that all such responding was of this form. Indeed, we observed some residual responding during the post-revaluation test that appeared to be impervious to outcome devaluation and therefore autonomous of the current incentive value, and we speculated that this responding was habitual and established by a process akin to the stimulus-response (S-R)/reinforcement mechanism embodied in Thorndike's classic *Law of Effect* (Thorndike, 1911). Research endorsing this dual-process theory has burgeoned in the intervening decades, and my purpose in the present chapter is to summarize and reflect on at least some these developments and thereby offer a tribute to my friend and mentor, Nick Mackintosh, in the form of an addendum to his 1994 volume.

I shall concentrate on three aspects of the subsequent research. The foundational work distinguishing between goal-directed and habitual behavior was exclusively conducted with rodents. However, it has become increasingly clear that exactly the same distinction applies to human instrumental behavior, which is my first topic. The second is the flourishing neurobiological research programs engendered by the dual-processes theory, but rather than attempting to articulate the neural circuits and mechanism mediating goal-directed and habitual behavior, I shall use this research to validate the basic distinction between the two forms of behavioral control. Finally, I shall conclude with a brief discussion of the theoretical accounts that have emerged during the last two decades. In this discussion, I shall retract my dismissal in 1994 of what I then called bidirectional theory, and now call ideomotor theory, as well as offering a brief description of reinforcement theory as an example of the computational theories that have been developed over the last two decades. A scholarly review of the extensive literature on action and habits that has been published since 1994 is beyond the scope of this chapter, and beyond my competence in the case of computational theories, and so I shall concentrate primarily on findings and ideas that accord with my own research and prejudices.

Of Rats and Humans

Although it has long been known that human instrumental responding, and the associated causal judgments, reflect variations in the action-outcome con-

tingency (Shanks & Dickinson, 1991), sensitivity to outcome revaluation had not been investigated in humans until John O'Doherty ask me to collaborate on a study run in his lab (Valentin, Dickinson, & O'Doherty, 2007). During instrumental training the human participants performed two responses, each of which produced a different fruit juice as an outcome. One of these outcomes was then devalued by allowing the participants to consume the juice to satiety before instrumental performance was assessed in the absence of the fruit juices. In accord with the rodent results, the human participants reduced responding that had, during training, produced the now-devalued outcome. Subsequent research, using clips from children's videos as outcomes and exposure to or specific satiety for one of the videos as the devaluation treatment, revealed that the capacity for goal-directed action develops by the age of 2 years (Klossek & Dickinson, 2012).

Human participants also manifest habitual behavior following instrumental training and, indeed, the conditions that favor habitual over goal-directed control in humans parallel those observed in rodent studies. Tricomi and colleagues (Tricomi, Balleine, & O'Doherty, 2009) replicated with humans Adam's (Adams, 1982) observation with rats that more extensive training can render performance resistant to outcome devaluation and therefore habitual, using symbolic food outcomes and consumption of the real food to satiety as the devaluation treatment. It is unlikely that the development of this behavioral autonomy of the current incentive value of the outcome is solely due to the fact a strong habit simply shortcuts goal-directed control. We trained 3-4 year-old children to perform one of two responses on each trial to see a video clip as an outcome, with clips from different videos being assigned to each response (Klossek, Yu, & Dickinson, 2011). For the single-action group, only one response option was available on each trial, whereas the children in the choice group chose between the two actions on each trial. On average, the choice group performed the two actions a similar number of times, which in turn closely matched the number of times the single-action group performed each response. As a result, the responses were reinforced with the same frequency both with- and between-groups and so should have had similar habit strengths. However, following devaluation of one of the videos by simple exposure or satiety, the two groups performed very differently in an extinction test with both response options. Whereas the single-action group behaved habitually, performing both responses equally frequently, the choice group preferential choose that action whose outcome had not been devalued.

Again this finding replicates the pattern previously observed with rats when trained with either a single-response or a choice procedure (Kosaki & Dickinson, 2010).

We argued that the variation in the development of behavioral autonomy arose from the different contingency experienced of the two groups. Once responding at a high and constant rate in the single-action condition after extended training, agents no longer experience the full causal contingency, specifically episodes in which they do not respond and do not receive the outcome. As a result, the action-outcome causal representation necessary for goal-directed action is not maintained. By contrast, during choice training the choice of one response and the receipt of its associated outcome also provides episodes in which the agent experiences that the alternative outcome does not occur in the absence of its response.Therefore, the children in the choice group continued to experience the full action-outcome causal contingency necessary for maintaining goal-directed control.

Not only does the onset of behavioral autonomy vary with the conditions of training but also with the conditions of testing. Using the Valentin et al. (2007) devaluation procedure, Schwabe and Wolf found that exposure to a social stressor prior to instrumental training rendered instrumental performance on test resistant to outcome devaluation (Schwabe & Wolf, 2009). Once again, an analogous effect has also been reported for rodents in that chronic-stress pre-exposure rendered lever pressing for a food outcome habitual and insensitive to contingency degradation (Dias-Ferreira et al., 2009). It is most likely that stress impacts on performance rather than learning because Schwabe and Wolf subsequently observed that goal-directed performance could be disrupted by administering the social stressor after training but prior to testing (Schwabe & Wolf, 2010).

Goal-directed performance appears particularly vulnerable. Hogarth and colleagues trained smokers on a tobacco-seeking response before devaluing the tobacco with health warnings and satiety (Hogarth, Field, & Rose, 2013). Tobacco-seeking was goal-directed in that the smokers refrained from performing this response on test unless an expectation of an alcoholic drink following the test had been induced just prior to the test. Yet again, this study was inspired by a rodent experiment in which goal-directed control over food seeking was disrupted by conducting the post-devaluation extinction test in an alcohol-paired context (Ostlund, Maidment, & Balleine, 2010). Furthermore, the Hogarth et al. study suggests that human drug seeking is

primarily under goal-directed control, a conclusion that again concurs with the rodent data. With a few exceptions (e.g., Miles, Everitt, & Dickinson, 2003), drug seeking by rats has been sensitive to revaluation of its outcome, the opportunity to take the drug (Hutcheson, Everitt, Robbins, & Dickinson, 2001; Olmstead, Lafond, Everitt, & Dickinson, 2001). However, in another study Hogarth's lab reported that among smokers the trait of impulsivity is negatively correlated with goal-directed control (Hogarth, Chase, & Baess, 2012).

Finally, a human research focus for dual-process theory is the obsessive-compulsive disorder (OCD) on the assumption that the habit process contributes to the compulsive behavior. Because the compulsions often appear to function as avoidance responses, Gillan and colleagues trained OCD patients and controls extensively on a shock avoidance paradigm before devaluing one of the two shock sources by instruction and removal of electrodes (Gillan et al., 2013). Although both groups showed a devaluation effect, the OCD patients exhibited more residual responding following devaluation, suggesting that the habitual process may make a greater contribution to avoidance in these patients. I did not discuss avoidance and negative reinforcement in the 1994 chapter because at the time there was only a single published study using a reinforcer revaluation procedure. In this report avoidance of a heat source was sensitive to revaluation produced by testing the rats in the cold (Hendersen & Graham, 1979). This evidence for the goal-directed status of rodent avoidance was recently confirmed by a study using a more conventional free-operant avoidance procedure in which the foot-shock reinforcer was revalued by presenting it non-contingently under either morphine or d-amphetamine (Fernando, Urcelay, Mar, Dickinson, & Robbins, 2014).

In summary, the last two decades have seen a proliferation of research using the outcome revaluation paradigm to determine the status of human action with the framework provided by dual-process theory. This research has yielded a remarkable concordance between human and rodent behavior, suggesting that the revaluation paradigm taps into fundamental and universal processes of behavioral control, at least in mammals. This conclusion is reinforced by the other major research theme since the 1994 chapter: the neurobiological investigation of the distinction between goal-directed and habitual behavior.

Neurobiological Dissociations

There is an inferential asymmetry in the assignment of control based on the outcome revaluation paradigm. If a reliable revaluation effect is observed (along with a sensitivity to the causal contingency), the target behavior can be characterized as goal-directed by definition. By contrast, identifying a response as habitual on the basis of a failure to detect a revaluation effect is more contentious because there could be a number of reasons why the revaluation treatment fails to change the incentive value of the outcome. Therefore, a prerequisite for inferring habitual control from the absence of a revaluation effect in an extinction test is independent evidence that the revaluation is effective within the test context. Often this evidence comes from a demonstration that the reinforcing properties of outcome are changed appropriately by the revaluation treatment when the outcome is presented contingent upon the action in the test context. Importantly, however, further evidence for the reality of the habit process comes from dissociations of outcome revaluation effects induced by neurobiological interventions. My primary aim in describing some of these dissociations is to evaluate whether they provide persuasive evidence for the dual-process theory at a psychological level rather than attempting to determine the neural basis of instrumental action.

Balleine and I (Balleine & Dickinson, 1998) reported the first demonstration of such a dissociation when we found that lesions of a rodent medial prefrontal cortex (PFC) structure, the prelimbic area, abolished sensitivity to both outcome revaluation and action-outcome contingency. This finding concurs with Valentin et al.'s (2007) observation of a greater fMRI bold signal in the ventromedial PFC when human participants performed a valued rather than devalued action on test. However, the role of PFC appears to be primarily in the acquisition rather than expression of goal-directed control in that prelimbic lesions are effective in abolishing sensitivity to outcome devaluation when given prior to training but not before testing (Ostlund & Balleine, 2005). A more likely site for the goal-directed engram is in the basal ganglia. Lesions of the posterior dorsomedial striatum (DMS) in rats, both prior to training and prior to testing, render instrumental performance insensitive to devaluation of the food outcome and contingency degradation (Yin, Ostlund, Knowlton, & Balleine, 2005).

These dysfunctions in goal-directed control are complemented by another set of lesions that interfere with the acquisition of behavioral autonomy to

yield a double dissociation between the two forms of control. Killcross and Coutureau found that lesions of another rat PFC structure, the infralimbic cortex, prevented the development of behavioral autonomy so that responding remained sensitive to outcome devaluation after a degree of training that induced habitual control in intact rats (Killcross & Coutureau, 2003). Moreover, in contrast to the selective role of the prelimbic PFC in acquisition of goal-directed control, temporary deactivation on the infralimbic PFC prior to testing reinstated goal-directed responding, suggesting the infralimbic cortex is involved in the deployment of habits. As in the case of goal-directed action, however, it is likely that habit, like goal-directed learning, takes place in the basal ganglia but in the lateral rather than medial dorsal striatum. Rats with pre-training lesions of the dorsolateral striatum remained sensitive to outcome devaluation (Yin, Knowlton, & Balleine, 2004) and action-outcome contingency degradation (Yin, Knowlton, & Balleine, 2006) after sufficient training to establish behavioral autonomy in intact animals. Once again, the human neurobiology of instrumental behavior reflects that of the rat. I have already noted that Tricomi et al. (2009) found that extensive training of humans on a button-press response for a food outcome led to habitual control. Importantly in the present context, concurrent fMRI scanning revealed that the development of behavioral autonomy was accompanied by activation of the (right) posterior putamen, a structure thought to be homologous to the rodent dorsolateral striatum.

In summary, goal-directed and habitual control have been doubly dissociated in two brain regions. In the PFC, lesions of the prelimbic and infralimbic areas disrupt goal-directed and habitual behavior, respectively, with the corresponding deficits being produced by (posterior) medial and lateral dysfunction in the dorsal striatum. These dissociations suggest that different neural circuits mediate the two forms of control, a conclusion that has recently received strong support from a diffusion tensor imagining (DTI) study of striatal white matter connectivity in humans by De Wit and colleagues. Goal-directed performance, as assessed by outcome devaluation, was positive correlated with the estimated tract strength between the ventromedial PFC and the caudate, a structure homologous to the rodent dorsomedial striatum (De Wit et al., 2012). In contrast, the putamen-premotor cortex connectivity was negative correlated with the outcome devaluation effect indicating that the stronger this connection, the greater the degree of habitual control.

Although this brief account gives only a limited description of the extensive neurobiological research on the dual-process theory over the last two decades since my 1994 chapter, it is sufficient to establish the reality of the dual control of instrumental behavior: the goal-directed and the habitual.

Theoretical Developments

As noted above, Adams and I attributed the residual responding follow outcome devaluation to habit learning through an S-R/reinforcement process. Although the intervening decades have seen the development of sophisticated computational theories of learning, the basic reinforcement account of habit learning has not been challenged, and the major theoretical developments have focused on goal-directed behavior.

Ideomotor Theory

Ideomotor theory has its origins in nineteenth century accounts of voluntary action (Stock & Stock, 2004) and, although the role of the ideomotor process was largely neglected during the twentieth century, the last decade has seen a renaissance of interest (Shin, Proctor, & Capaldi, 2010). When applied to instrumental learning, ideomotor theory argues that such learning leads to the formation of two associations. The first is a Pavlovian S → O association brought about by pairings of the outcome (O) with the stimulus context (S) in which the instrumental training takes place, whereas the second is the O → R association generated by the instrumental contingency between the response (R) and outcome. Amalgamating these two associations enables the stimulus to activate the response through an S → O → R chain.

Although the theory assumes that the associations are formed concurrently during instrumental training, the most compelling evidence for the ideomotor account comes from studies of outcome-specific Pavlovian-instrumental transfer (PIT) in which the S → O and O → R associations are trained separately. The integration of the two associations to yield a S → O → R chain is then assessed by the capacity of the stimulus to elicit the response even though the response has never been previously trained in the presence of the stimulus. For example, Watson and colleagues trained human

participants to press, for instance, a right-hand (RH) key for chocolate and left-hand (LH) key for popcorn (Watson, Wiers, Hommel, & De Wit, 2014). Following this instrumental training, the participants received Pavlovian pairings of one abstract stimulus (Sc) for chocolate and another stimulus (Sp) for popcorn. According to ideomotor theory, two associative chains, Sc → chocolate → RH press and Sp → popcorn → LH press, should have been established by this training and, in accord with this prediction, the participants chose to press the RH key in the presence of Sc and the LH in the presence of Sp when for the first time the response options were available in the presence of the stimuli.

In my 1994 chapter, I considered ideomotor theory in the guise of bidirectional theory, which assumes that associations are bidirectional and was the account of instrumental learning espoused by Pavlov and his students (Asratyan, 1974). However, I dismissed the theory for two reasons. First, the typical PIT procedure trains each response in separate sessions with the result that the reinforcer could have functioned, not only as an outcome of its response, but also as a stimulus for the response as the response was reinforced in a context in which the outcome was presented. Indeed, the response would also have been reinforced in presence of an activated representation of the outcome in the training context. Consequently, following such training, this form of PIT could simply have been due to the fact the stimulus reinstated the stimulus training conditions for the response and therefore is compatible with a S-R/reinforcement theory of instrumental conditioning. Importantly, however, this account cannot be applied to the transfer observed by Watson and colleagues (2014) because they trained the LH and RH key presses concurrently in the same stimulus context with the consequence that each response should have been reinforced in the presence of activated representations of both outcomes. With this training regime the transfer must have been mediated by O → R associations generated by the instrumental response-outcome contingencies, thereby rendering the S → O → R ideomotor chain a plausible component of goal-directed action.

Integrating Ideomotor and Associative-Cybernetic Mechanisms

My second concern with the bidirectional or ideomotor theory is the absence of any mechanism by which the incentive value of the outcome can modulate

behavior appropriately. Consider a punishment contingency in which an action yields an aversive event, such as a shock, rather than a reward. Under such contingency, the stimulus situation should excite a representation of the shock, which in turn should activate rather than suppress the response according to simple ideomotor theory. Furthermore, the failure of motivational manipulations to impact on outcome-specific PIT reinforces the fact the ideomotor mechanism fails to capture the motivation control of goal-directed behavior. For example, Watson and colleagues (2014) found that the magnitude of the PIT effect was unaffected by devaluing one of the outcomes by sating the participants on this food immediately prior to the transfer test even though the same manipulation produced a standard outcome devaluation effect. This insensitivity of outcome-specific PIT to change in the incentive value of the outcome accords with the results of a number of rodent studies (Rescorla, 1994) and suggests that the ideomotor outcome representation encodes the sensory features of the outcome but not its incentive value.

In response to these findings, De Wit and I (De Wit & Dickinson, 2009, 2016) have suggested that the ideomotor mechanism can be integrated into an associative-cybernetic (A-C) model of goal-directed and habitual behavior, the theory I favored in the 1994 chapter. Although I shall not reiterate the details of the model here, the core idea is that the incentive value of the outcome is evaluated at the same time that potential instrumental responses are primed. An excitatory influence generated by this evaluation is then applied to any primed response with the most strongly primed response being executed. In the 1994 statement of the model, the response priming originated through in a stimulus-response habit mechanism, and what De Wit and I subsequently suggested is that this source of priming is supplemented by the ideomotor mechanism. Therefore, activating a sensory representation of an outcome primes its associated response through the ideomotor association in parallel with a motivational evaluation of the current incentive value of this outcome which, if positive, feedbacks to the response mechanism to cause the execution of the primed instrumental response.

In my 1994 chapter, I also considered another form of transfer, general PIT, that in contrast to the outcome-specific form is sensitive to motivational manipulations and which subsequent research has shown is dissociable from outcome-specific PIT. Corbit and Balleine developed a procedure for demonstrating the two forms of PIT in the same study. Instead of just training two Pavlovian stimuli, each associative with one of the two instrumental

outcomes, they also trained a third stimulus, which predicted a food outcome that differed from both instrumental outcomes. Consequently, presenting this third stimulus while the animals were engaged in instrumental responding should reveal any general transfer that is not mediated by an outcome common to the Pavlovian and instrumental training. In accord with the finding that outcome-specific PIT is unaffected by motivational manipulations, sating the rats with their maintenance diet before testing had no effect on this form of transfer but reduced general PIT (Corbit, Janak, & Balleine, 2007). Moreover, dysfunctions within the basolateral amygdala and shell of the accumbens impacted on outcome-specific PIT, whereas disruption of the central amygdala and core of the accumbens attenuated the general form (Corbit & Balleine, 2005; Corbit & Balleine, 2011). Furthermore, the dissociation within the rodent amygdala accords with the pattern of fMRI activation seen with the human amygdala during the two forms of transfer (Prévost, Liljeholm, Tyszka, & O'Doherty, 2012)

General PIT provides the mechanism for motivating habits rather than goal-directed actions. Many years ago, we found that interval schedules more readily establish habitual control than do ratio schedules (Dickinson, Nicholas, & Adams, 1983), and Wiltgen and colleagues have recently exploited this distinction to demonstrate greater general PIT on interval-trained habitual responding than on ratio-trained goal-directed behavior in mice (Wiltgen et al., 2012). Therefore, habits could be motivated by Pavlovian conditioning to the training context induced by the instrumental outcome or reinforcer.

Reinforcement Theory

Both the ideomotor and associative-cybernetic theories are expressed in terms of relatively simple associative psychological structures and mechanisms. In contrast, the last two decades have seen the emergence of plethora of computational-based theories of goal-directed action and especially human choice behavior (Dolan & Dayan, 2013; Solway & Botvinick, 2012), As the sophistication of these theories lies beyond the scope of my expertise, I can offer no more than a brief descriptive account of the most influential of these theories, reinforcement learning theory (Sutton & Barto, 1998; Dolan & Dayan, 2013), which has origins in machine learning and presents a normative account of both goal-directed and habitual instrumental action.

The theory analyses instrumental contingencies into *states* (stimuli), *actions*, *transitions* between states produced by the actions (instrumental contingencies) and *utilities* (incentive values) and assumes that the agent will learn or decide upon a *policy* that specifies the action for each state that optimizes the gained utilities in the long run. Within this framework, there are two ways of deriving a policy. Goal-directed behavior is controlled by *model-based* computations in which the agent engages in prospective planning using a learned decision tree of the state transitions produced by actions that models the instrumental contingencies in the environment. Starting with the current state, the policy with the highest utility is determined by using mental simulation to search the decision tree.

This model-based control contrasts with a less computationally demanding *model-free* control. This type of control uses reward prediction-error learning to link the policy of performing a particular action in a given state with a summary of the utilities of the subsequent states that have followed this action in the past. The probability of adopting this policy is thus determined by the sum of the past utilities that are associated with the policy. Model-free control mediates habitual behavior because the utility of a policy is divorced from the identity of the particular states that have grounded that the summed utility in the past so that the policy is not directly sensitive to changes in the current utility of a particular state through outcome revaluation. To adapt to such revaluation, the model-free system has to be retrained with the changed utilities just as habits have to adjust to alterations in reinforcing properties of revalued outcome through the S-R/reinforcement mechanism.

There are certain similarities between the reinforcement learning and A-C accounts of instrumental behavior. The associative representations of the instrumental response-outcome contingences in the A-C model is analogous to a simple action-state model of reinforcement theory. The priming of a potential response via response → outcome association in the A-C theory is akin to a policy yielded by a search of the model from the current state to the outcome state in reinforcement theory, whereas the priming of a response by an ideomotor outcome → response association could be matched to searching the model from an outcome or goal state to the current state. Moreover, both theories assume that learning in the model-based and model-free processes occur concurrently during instrumental training.

Where they differ, however, is in their account of the interaction between the two processes at output. Because the feedback loop of A-C theory through

the representation of the outcome and its motivation evaluation acts by modulating the output of the same response mechanism as that mediating habits, the output of each system is assumed to summate in controlling responding (Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995). By contrast, Daw and colleagues suggested that the arbitration between the model-based and model-free controllers is based on the uncertainty of the utilities produced by each process with the selected controller being that which yields the most certain prediction (Daw, Niv, & Dayan, 2005). Moreover, they demonstrated by simulation that this arbitration process can predict both the development of behavioral autonomy with extensive training of a single response (e.g. Adams, 1982) and the maintained sensitivity to outcome devaluation following extensive training of at least two response-outcome contingencies (e.g., Klossek et al., 2011; Kosaki & Dickinson, 2010).

Beyond Instrumental Conditioning

As this brief summary demonstrates, the dual-process theory of instrumental conditioning has been an active and growing research topic during the decades since my 1994 chapter was published. The neurobiological dissociations discovered during the last two decades confirm the reality of the distinction between goal-directed and habitual behavior, which have been established as fundamental forms of behavioral control in both humans and other animals. I shall now conclude this addendum with an important distinction that has emerged since the 1994 chapter, that between goal-directed behavior and future planning (Dickinson, 2011), which will provide an important research focus for the next decade or so.

The definition of goal-directed behavior endorsed at the outset of this chapter requires that performance is sensitive to the current incentive value of the outcome and therefore to the agent's current motivational state. However, in the case of human behavior at least, there are numerous anecdotal examples of actions that we perform in the service of future rather than current incentive values. Whereas I might go to the refrigerator under goal-directed control to get the ingredients for my lunch in the service of my current hunger, I also go to the supermarket at a time when I am not hungry to buy the food for tomorrow's lunch. The distinction between the control of action by future rather than current motivational states can be illustrated by

the case of food caching by animals that store food in a time of plenty in order to have a supply of food at a time of scarcity in the future.

Clayton and colleagues (Cheke & Clayton, 2012; Correia, Dickinson, & Clayton, 2007) used a food-caching paradigm with jays and varied the relative incentives values of two types of food at caching in the morning and at recovery of the caches in the afternoon. The incentive values were manipulated by pre-feeding one of the foods to reduce its value through specific satiety. The procedure contrasted the control of caching by the incentive value of a food at the time of caching with that at the time of recovery. On the first day, the birds cached more of the non-prefed, and therefore more valuable food, a choice that reflected the relative incentive values at the time of caching. However, the birds were then pre-fed the other food just prior to recovery in the afternoon, thereby dissociating the relative values of the two foods at caching and recovery. One food was valuable at the time of caching and the other at the time of recovery. At issue, then, was which food would the jays choose to cache on the second day? The fact that they switched their preference to caching the food that had been valuable at recovery on the previous day rather than the one that was valuable at the time of caching demonstrates that they are capable of acting in the service of a future need. Caching was therefore controlled by the incentive value at the time of recovery rather than the value at the time of caching (Cheke & Clayton, 2012).

The choice of what to cache appears to be instrumental in that it is controlled by the contingency between the action, caching a particular food, and the outcome, the recovery of that food. However, the control of caching by a future motivational state is problematic for the theories of goal-directed behavior that we have considered for a number of reasons. First, to the extent that the theory requires an association between a response and outcome, the time interval between these events is beyond that typical of standards forms of associative learning. In response to this issue, I suggested that the recovery of a particular food retrieves a memory of caching that food, allowing its recovery to be associated with its caching and thereby bridging the long interval (Dickinson, 2011). There is good evidence that the jays remember the caching episode at the time of recovery (Clayton & Dickinson, 1998).

This mnemonic-associative theory still leaves the problem of why the jays cached the prefed food even though this had a lower incentive value than the

non-prefed food as indicated by the fact that they eat more of this non-prefed food at the time of caching. Food caching is, of course, an adaptive specialization, and so one possibility is that it calls upon specialized cognitive resources, which are not available for the control of general goal-directed behavior. Alternatively, as Suddendorj and Corballis have suggested in the case of human behavior, future planning deploys general cognitive processes that support what they call *mental time travel* (Suddendorf & Corballis, 1997), which generates action in the service of future needs. Whatever the merits of this account, the relationship between standard goal-directed behavior under the control of current incentive values and actions directed at future incentive values will feature in any further addendum written in a decade or so time.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, *34B*(2), 77-98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *33B*, 109-122.

Asratyan, E. A. (1974). Conditioned reflex theory and motivational behaviour. *Acta Neurobiologiae Experimentalis*, *43*, 15-31.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407-419.

Cheke, L. G., & Clayton, N. S. (2012). Eurasian jays (*Garrulus glandarius)* overcome their current desires to anticipate two distinct future needs and plan for them appropriately. *Biology Letters*, *8*(2), 171-175.

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub-jays. *Nature*, *395*, 272-274.

Corbit, L. H., & Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. *Journal of Neuroscience*, *25*, 962-970.

Corbit, L. H., & Balleine, B. W. (2011). The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *Journal of Neuroscience*, *31*(33), 11786-11794.

Corbit, L. H., Janak, P. H., & Balleine, B. W. (2007). General and outcome-specific forms of Pavlovian-instrumental transfer: The effect of shifts in motivational

state and inactivation of the ventral tegmental area. *European Journal of Neuroscience, 26*, 3141-3149.

Correia, S. P. C., Dickinson, A., & Clayton, N. S. (2007). Western scrub-jays anticipate future needs independently of their current motivational state. *Current Biology, 17*, 856-861.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*, 1704-1711.

De Wit, S., & Dickinson, A. (2009). Associative theories of goal-directed behaviour: A case for animal–human translational models. *Psychological Research, 73*(4), 463-476.

De Wit, S., & Dickinson, A. (2016). Ideomotor mechanism of goal-directed behavior. In T. S. Braver (Ed.), *Motivation and cognitive control* (pp. 123-142). New York: Psychology Press.

De Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., van de Vijver, I., & Ridderinkhof, K. R. (2012). Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. *Journal of Neuroscience, 32*(35), 12066-12075.

Dias-Ferreira, E., Sousa, J. C., Melo, I., Morgado, P., Mesquita, A. R., Cerqueira, J. J., & de Sousa, N. (2009). Chronic stress causes frontostriatal reorganization and affects decision-making. *Science, 31*(July), 621-625.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society (London), B, 308*, 67-78.

Dickinson, A. (1994). Instrumental conditioning. In N. J. Mackintosh (Ed.), *Animal learning and cognition* (pp. 45-79). San Diego, CA: Academic Press.

Dickinson, A. (2011). Goal-directed behavior and future planning in animals. In R. Menzel & J. Fischer (Eds.), *Animal thinking: contemporary issues in comparative cognition* (pp. 79-91). Cambridge, MA: MIT Press.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning and Behavior, 23*, 197-206.

Dickinson, A., & Mulatero, C. W. (1989). Reinforcer specificity of the suppression of instrumental performance on a non-contingent schedule. *Behaviour Processes, 19*(1-3), 167-180.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptiblity to reinforcer devaluation. *Quarterly Journal of Experimental Psychology, 35B*, 35-51.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*, 312-325.

Fernando, A., Urcelay, G., Mar, A., Dickinson, A., & Robbins, T. (2014). Free-operant avoidance behavior by rats after reinforcer revaluation using opioid agonists and d-amphetamine. *Journal of Neuroscience, 34*(18), 6286-6293.

Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., . . . Robbins, T. W. (2013). Enhanced avoidance habits in obsessive-compulsive disorder. *Biological Psychiatry*, *75*, 631-638.

Hendersen, R. W., & Graham, J. (1979). Avoidance of heat by rats: Effects of thermal context on the rapidity of extinction. *Learning and Motivation*, *10*, 351-363.

Hogarth, L., Chase, H. W., & Baess, K. (2012). Impaired goal-directed behavioural control in human impulsivity. *The Quarterly Journal of Experimental Psychology*, *65*(2), 305-316.

Hogarth, L., Field, M., & Rose, A. K. (2013). Phasic transition from goal-directed to habitual control over drug-seeking produced by conflicting reinforcer expectancy. *Addiction Biology*, *18*(1), 88-97.

Hutcheson, D. M., Everitt, B. J., Robbins, T. W., & Dickinson, A. (2001). The role of withdrawal in heroin addiction: Enhances reward or promotes avoidance? *Nature Neuroscience*, *4*, 943-947.

Killcross, A. S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*, 400-408.

Klossek, U. M. H., & Dickinson, A. (2012). Rational action selection in 1 1/2-to 3-year-olds following an extended training experience. *Journal of Experimental Child Psychology*, *111*, 197-211.

Klossek, U. M. H., Yu, S., & Dickinson, A. (2011). Choice and goal-directed behavior in preschool children. *Learning and Behavior*, *39*, 350-357.

Kosaki, Y., & Dickinson, A. (2010). Choice and contingency in the development of behavioral autonomy during instrumental conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(3), 334-342.

Miles, F. J., Everitt, B. J., & Dickinson, A. (2003). Oral cocaine seeking by rats: Action or habit? *Behavioral Neuroscience*, *117*, 927-938.

Olmstead, M. C., Lafond, M. V., Everitt, B. J., & Dickinson, A. (2001). Cocaine-seeking by rats is a goal directed action. *Behavioral Neuroscience*, *115*, 394-402.

Ostlund, S. B., & Balleine, B. W. (2005). Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *Journal of Neuroscience*, *25*, 7763-7770.

Ostlund, S. B., Maidment, N. T., & Balleine, B. W. (2010). Alcohol-paired contextual cues produce an immediate and selective loss of goal-directed action in rats. *Frontiers in Integrative Neuroscience*, *4*(19).

Prévost, C., Liljeholm, M., Tyszka, J. M., & O'Doherty, J. P. (2012). Neural correlates of specific and general Pavlovian-to-instrumental transfer within human amygdalar subregions: A high-resolution fMRI study. *Journal of Neuroscience*, *32*(24), 8383-8390.

Rescorla, R. A. (1994). Transfer of instrumental control mediated by a devalued outcome. *Animal Learning and Behavior*, *22*, 27-33.

Schwabe, L., & Wolf, O. T. (2009). Stress prompts habit behavior in humans. *Journal of Neuroscience*, *29*(22), 7191-7198.

Schwabe, L., & Wolf, O. T. (2010). Socially evaluated cold pressor stress after instrumental learning favors habits over goal-directed action. *Psychoneuroendocrinology*, *35*(7), 977-986.

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory and Cognition*, *19*, 353-360.

Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychological Bulletin*, *136*(6), 943-947.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120-154.

Stock, A., & Stock, C. (2004). A short history of ideo-motor action. *Psychological Research*, *68*, 176-188.

Suddendorf, T., & Corballis, M. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs*, *123*, 133-167.

Thorndike, E. L. (1911). *Animal intelligence: experimental studies.* New York: Macmillan.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*, 2225-2232.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, *27*, 4019-4026.

Watson, P., Wiers, R. W., Hommel, B., & De Wit, S. (2014). Working for food you don't desire. Cues interfere with goal-directed food-seeking. *Appetite*, *79*, 139-148.

Wiltgen, B. J., Sinclair, C., Lane, C., Barrows, F., Molina, M., & Chabanon-Hicks, C. (2012). The effect of ratio and interval training on Pavlovian-instrumental transfer in mice. *PLoS ONE*, *7*(10), e48227.

Yin, H. H., Knowlton, B. J., & Balleine, B. B. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*, 181-189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2006). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. *Behavioural Brain Research*, *166*, 189-196.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*, 513-523.

*Asymmetry in the Discrimination of Auditory Intensity: Implications for Theories of Stimulus Generalisation**

Richard A. Inman, Robert C. Honey,
John M. Pearce
Cardiff University, UK

Abstract. In each of four experiments, rats received a magnitude discrimination between two intensities of the same clicker. The trials with the clicker were separated by an intertrial interval (ITI) in Experiment 1, and it was found that the discrimination was acquired more readily if the delivery of food was signalled by the loud rather than the soft clicker, than when this relationship was reversed. The remaining experiments explored the possibility that the asymmetry was a consequence of inhibition associated with the ITI generalising more to the trials with the soft than the loud clicker. In contradiction to this proposal, the asymmetry observed in Experiment 1 was also found if the trials were not separated by an ITI (Experiments 2 and 3), and if the cues present during the ITI included a clicker that was louder than the training stimuli (Experiment 4). The results are explained with a modified version of the configural theory of discrimination learning proposed by Pearce (1994).

Introduction

In his book, *The Psychology of Animal Learning*, Mackintosh (1974) outlined an account of how discriminations are solved that is still of considerable influence. Our purpose in the present chapter is to evaluate this analysis by exploring how well it explains the outcome of discriminations based on stimulus magnitude. The account favoured by Mackintosh can be introduced with the following quotation, which summarises the assumptions on which it is based.

* This research was supported in part by a grant from the United Kingdom Biotechnology and Biological Sciences Research Council (BB/H006176).

> These assumptions are that discrimination training can be reduced to the scheduling of differential reinforcement in the presence of S+ and S–; that this differential reinforcement establishes S+ as an excitatory stimulus and S– as an inhibitory stimulus; that the excitatory and inhibitory processes established to S+ and S– generalize to other, similar stimuli (including each other); and finally, that discriminative performance is a consequence of this interaction of excitatory and inhibitory gradients.[1]

Mackintosh (1974) then went on to conclude that: "Among the several consequences of this set of assumptions, it follows that the rate of discrimination learning will be inversely related to the degree of similarity between S+ and S–" (p. 532).

If it is accepted that the similarity between two stimuli is symmetrical, then it follows from the foregoing analysis that a discrimination between them will be mastered just as readily, regardless of which one serves the role of S+. In other words, a discrimination between any pair of stimuli will always be symmetrical. An apparent challenge to this conclusion, and thus a challenge to the assumptions on which it is based, is provided by discriminations involving different magnitudes of the same stimulus. Typically, such magnitude discriminations reveal an asymmetry, where acquisition is more rapid when the high magnitude rather than the low magnitude stimulus signals reward. Assuming that the similarity between a low and high magnitude stimulus is also symmetrical, then it follows that a discrimination between them will also be symmetrical.

The asymmetry in magnitude discriminations is widespread. It has been revealed with the magnitudes of quantity (Inman, Honey, Eccles, & Pearce, 2016; Inman, Honey, & Pearce, 2015; Vonk, 2012; Watanabe, 1997); temporal duration (e.g. Bouton & García-Gutiérrez, 2006; Bouton & Hendrix, 2011; Kyd, Pearce, Haselgrove, Amin, & Aggleton, 2007; Todd, Winterbauer, & Bouton, 2010); odour intensity (Pelz, Gerber, & Menzel, 1997); intensity of auditory stimuli (Jakubowska & Zielinski, 1976; Pierrel, Sherman, Blue, & Hegge, 1970); and the length of wall panels in a rectangular arena (Kosaki, Jones, & Pearce, 2013). All but one of the foregoing results was reported after the publication of Mackintosh (1974) and it thus not surprising that he

1. Mackintosh, 1974, p. 532.

did not offer an explanation for the remarkably consistent pattern of results they reveal. Nonetheless, following from proposals put forward by Perkins (1953), and Logan (1954), the next quotation points to one way in which the asymmetry might be reconciled with his theoretical account of discrimination learning.

"The greater the intensity of a CS, the less the generalization of inhibition from non-reinforced background stimuli and therefore the faster the rate of learning" (Mackintosh, 1974, pp. 532-3).

Consider an experiment by Inman et al. (2016), in which rats were required to solve a successive discrimination between two quantities of black squares (5 or 20) presented on a white screen that was also white during the interval between each trial. From the above proposals, the white screen by itself can be expected to acquire inhibition through its non-reinforced exposure during the intertrial interval (ITI), which will generalise to a greater extent to the low quantity than the high quantity stimulus. This asymmetry in generalisation will then hinder excitatory conditioning with the low quantity stimulus to a greater extent than the high quantity stimulus and result in the 5+ 20– discrimination being acquired more slowly than 20+ 5– (see also Moore, 1964).

In order to evaluate the foregoing analysis, Inman et al. (2016) tested two predictions that follow directly from it. One prediction is that the asymmetry will not be observed if the stimuli that differ in magnitude are presented immediately after each other, rather than separated by an ITI. This manipulation will remove the source of inhibition responsible for the asymmetry, and permit the discriminations to progress at the same rate, regardless of which magnitude signals the reinforcer. This prediction was tested with rats using the quantity discrimination just described. For two groups trained with an ITI, during which the white screen was presented without any black squares, the usual asymmetry was observed: the large+ small– discrimination was acquired more readily than small+ large–. In stark contrast, and in support of the above prediction, this asymmetry was not observed when presentations of the experimental cues were not separated by an ITI.

The second prediction is that it should be possible to reverse the asymmetry observed with magnitude discriminations, if the interval between successive trials is filled with more intense stimulation than that provided by the larger of the two training stimuli. The non-reinforced exposure to the ITI will enable the cues present during this interval to enter into an inhibitory association, but now the generalisation of this inhibition will be greater to

the larger than the smaller of the two experimental stimuli. The discrimination in which the large stimulus signals the reinforcer will then be disrupted to a greater extent than if the small stimulus serves this purpose.

To test this prediction, Inman et al. (2016) again used a quantity discrimination involving different numbers of small black squares presented on a white screen. During the interval between successive trials the white screen was covered in a larger number of black squares than when the experimental stimuli were displayed. This treatment resulted in an asymmetry in the acquisition of the magnitude discriminations, but it was opposite to that seen when no squares were shown on the screen during the ITI. The findings from the two experiments just described have also been obtained with pigeons (Inman et al., 2015).

The results from both experiments thus imply that the asymmetries found with magnitude discriminations do not challenge the theoretical proposals of Mackintosh (1974) and, in fact, are entirely consistent with them. For the sake of theoretical elegance and parsimony, it is tempting to go one step further and suggest that this explanation will apply to the asymmetry that is found with any discrimination based on stimulus magnitude, and not just quantity. There is, however, very little evidence by which this suggestion can be judged. Accordingly, the purpose of the experiments described below was to determine whether the explanation for the asymmetry that has been found with discriminations based on quantity also applies to discriminations based on a different kind of magnitude — auditory intensity. Experiment 1 was conducted in order to confirm that there is an asymmetry in the acquisition of a discrimination based on the intensity of a clicker, with the clicker being silent during the interval between successive presentations of the experimental stimuli. The next three experiments then examined the effects on the acquisition of the intensity discriminations of removing the ITI (Experiments 2 and 3), and of presenting a clicker of higher intensity than the experimental stimuli during the ITI (Experiment 4). In contrast to the experiments involving different numbers of black squares, these manipulations had rather little impact on the manner in which the discrimination was solved. An implication of this finding is that the asymmetry in discriminations based on magnitude may not always occur for the same reason, and that at least on some occasions this effect may pose a challenge to a well-established account of discrimination learning (Mackintosh, 1974). The significance of this conclusion is pursued in the final section of this chapter.

Experiment 1

One group of ten and one group of nine rats received appetitive Pavlovian conditioning in dark chambers with a clicker presented through a speaker in the roof. The intensity of the clicker was either soft, 78 dB, or loud, 87 dB. For one of these intensities, S+, trials always terminated with the delivery of sucrose solution into a food well. The solution could be reached by poking the snout through a hole in the chamber wall. For the other intensity stimulus, S–, trials were never followed by a delivery of sucrose. Over the course of training it was anticipated that rats would make more snout entries during S+ than S–.

For the loud+ group the presentation of the loud clicker served as S+ and the soft clicker served as S–. The stimuli were presented for 15 sec at a time and each trial was separated by an ITI (mean duration 6 min, range 4-8 min). For the soft+ group, the soft clicker served as S+ while the loud clicker was S–. If the asymmetry identified in other discriminations of auditory intensity (e.g. Jakubowska & Zielinski, 1976, Pierrel et al., 1970) is reliable, it follows that the loud+ group will solve its discrimination more rapidly than the soft+ group.

The mean rates of snout entry during S+ and S- for each of the 14 sessions of training are shown for the loud+ group in the upper left-hand panel of figure 1, and the equivalent results for the soft+ group can be seen in the upper-right hand panel of the same figure. The figures also show the mean rates of responding that were recorded for the two groups during 15 sec intervals prior to each trial. The experiment replicated the asymmetry that has been found with other discriminations based on stimulus intensity, with the loud+ soft- discrimination being acquired more readily than soft+ loud–.

In order to compare the performance of the two groups, their results were transformed to discrimination ratios of the form $A/(A + B)$, where A and B were the mean rates of responding during reinforced and non-reinforced trials, respectively. The ratios can be seen in the lower panel of figure 1, which shows that the ratios were generally higher for the loud+ group than the soft+ group. In support of this observation, a two-way, Group × Session ANOVA of the ratios revealed a significant main effect of group, $F(1,18) = 10.13$, $p = .005$, $\eta_p^2 = .36$. In addition, there was a significant effect of session, $F(13,234) = 3.30$, $p < .001$, $\eta_p^2 = .15$, but the interaction between these factors was not significant, $F < 1$.

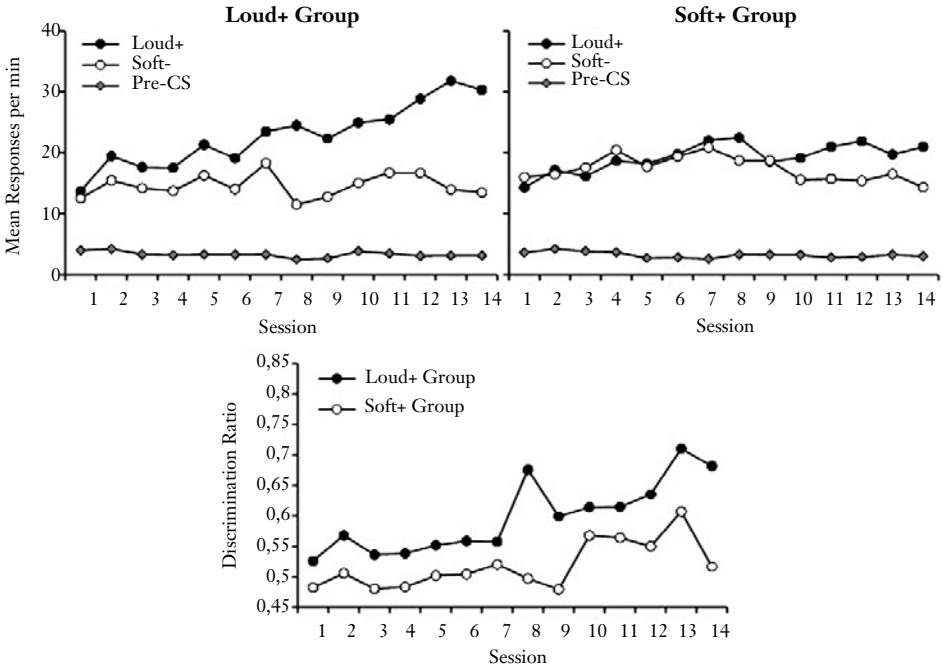

FIGURE 1. The mean rates of responding to S+ and S– for the 14 sessions of training for the loud+ group (top left panel) and the soft+ group (top right panel) of Experiment 1, and the discrimination ratios for both groups (lower panel).

Having established that there is an asymmetry in the acquisition of a discrimination based on different intensities of a clicker, the next two experiments explored how this outcome is affected by removing the interval between successive presentations of this cue. If the asymmetry is a result of generalisation from non-reinforced cues present during the ITI then removing this interval will also remove the asymmetry.

EXPERIMENT 2

An obvious way to investigate the effect of removing the ITI on the asymmetry that was found in Experiment 1, would be to repeat the experiment but with no interval between the sequence of 15 sec trials with the loud and soft clicker. A preliminary investigation revealed, however, that this method of training did not result in the successful acquisition of the discrimination.

The details of the experimental design were therefore based on the successful methodology adopted by Inman et al. (2016), who investigated how the removal of the ITI affects the acquisition of a magnitude discrimination based on number. Four groups of eight rats received eight sessions in which they were trained to discriminate between a loud, 82-dB, and a soft, 57-dB, clicker. The duration of each reinforced trial was 73 sec, during which three pellets of food were delivered individually into a food well at random intervals. For the loud+/ITI and the loud+/no-ITI groups, the loud clicker served as S+ and the soft clicker served as S–, whereas for the soft+/ITI and the soft+/no-ITI groups, S+ was the soft clicker and S– the loud clicker. For the two groups trained with an ITI, the loud and soft clicker were presented in alternation and separated by an interval of 73 sec. For the two groups trained without an ITI this interval was 0 sec. In order to equate the duration of the session across groups, whilst keeping the same number of trials, each non-reinforced trial with a clicker was 73 sec for the groups trained with an ITI, and 219 sec for the groups trained without an ITI.

The mean rates of responding prior to the delivery of the first food pellet for each reinforced trial, and during an equivalent period for each non-reinforced trial, for the two groups trained with an ITI can be seen in the upper panel of figure 2. In keeping with the previous experiment, the loud+ soft– discrimination was acquired more readily than soft+ loud–. The results displayed in the lower row of figure 2 demonstrate that a similar asymmetry was observed with the two groups trained without an ITI. Discrimination ratios calculated in the same manner as for Experiment 1 are presented for the four groups in figure 3. The ratios were generally larger for the groups trained with a loud clicker as S+, than for the groups trained with a soft clicker as S+, when training was conducted with, and without an ITI. In support of this observation, a three-way ANOVA with the between-group factors of ITI (present or absent), and intensity (whether the loud or soft stimulus signalled food), and the within-group factor of session revealed a significant three-way interaction, $F(7,196) = 2.54$, $p = .016$, $\eta_p^2 = .08$. Subsequent tests of simple-main effects revealed that the discrimination ratios were significantly larger for the groups trained with the loud, rather than the soft clicker as the signal for food, when training took place with an ITI, on Sessions 5, 6 and 7, and when training took place without an ITI, on Sessions 1 and 2, Fs$(1,224) > 7.73$, ps $< .01$, $\eta_p^2 > .03$.

Inspection of the right-hand panel of figure 3 shows a difference between the two No-ITI groups at the outset of testing. This difference raises the

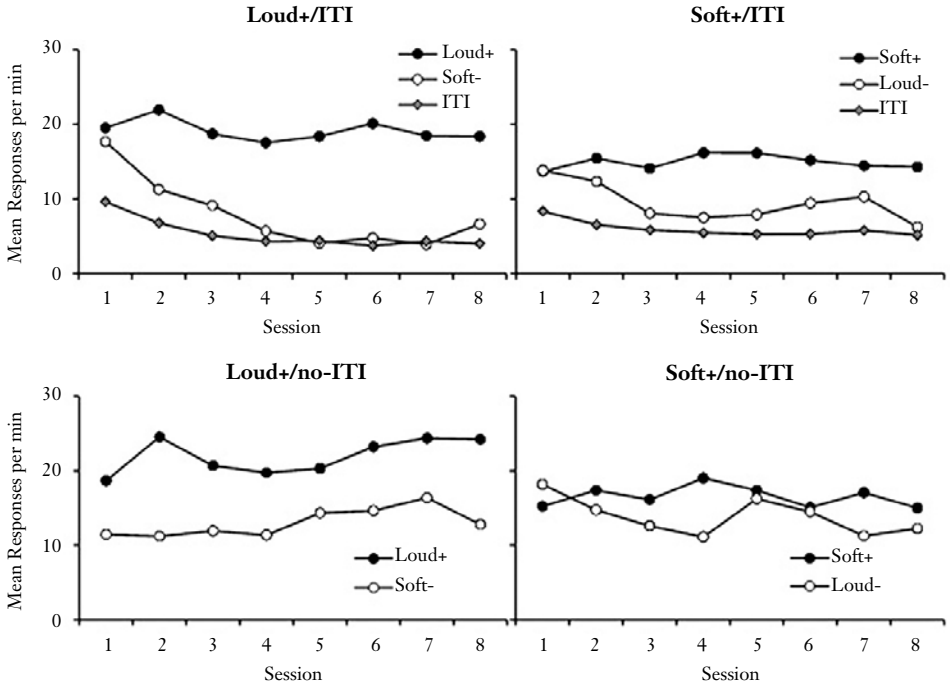

FIGURE 2. The mean rates of responding to S+ and S– for the eight sessions of training in Experiments 2 for the loud+/ITI group (top left panel), the soft+/ITI group (top right panel), the loud+/no-ITI group (bottom left panel), and the soft+/no-ITI group (bottom right panel).

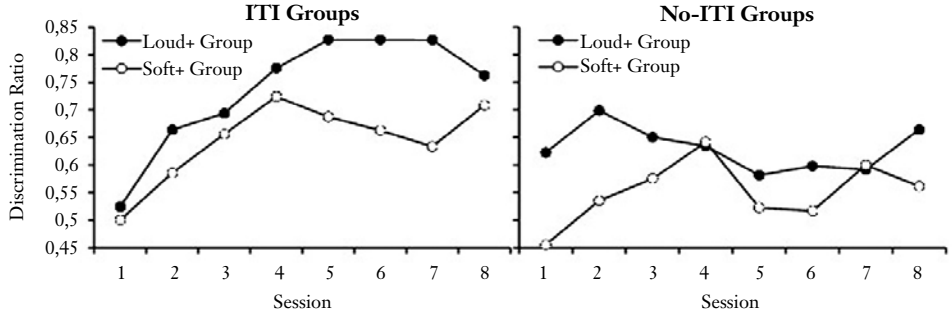

FIGURE 3. Discrimination ratios for each session of Experiment 2 for groups trained with an ITI (left-hand panel) and for the groups without an ITI (right-hand panel).

possibility that the asymmetry was due to an unconditioned influence of stimulus intensity that energised responding to a greater extent during the loud than the soft clicker. Examination of the results from the first trial with S+ and S– do not support this prediction. For the loud+/No ITI group the mean number of responses per minute was 11 on the first trial with the loud clicker, and 15 on the first trial with the soft clicker, which is opposite to the pattern just suggested. In the case the soft+/ITI group, the equivalent values were 10 and 11 for the loud clicker and soft clicker, respectively which, once again, does not support the explanation just presented. Neither of these differences was significant, $ts(7) \leq .34$, $ps > .10$.

In view of the finding by Inman et al. (2016), where removing the ITI from a quantity discrimination removed the asymmetry observed with an ITI, the present results came as something of a surprise. Despite the removal of the ITI, the discrimination with the loud clicker signalling food was acquired more readily than when the soft clicker signalled food. It is worth noting that apart from the use of different stimuli, the details of the present experiment were much the same as those used for the experiment by Inman et al. (2016). The failure to replicate the earlier finding is thus unlikely to be due to some procedural difference between the two experiments. Instead, it would seem that different explanations are required for the asymmetry that is found with discriminations based on the magnitudes of quantity and intensity. Before pursuing the implications of this conclusion, the next two experiments were conducted in order to lend it further weight.

Experiment 3

An unusual aspect of the design of the previous experiment is that the amount of exposure to S– was three times that to S+ in the no ITI condition, whereas when the ITI was present the amount of exposure to S+ and S– was the same. Although this feature of the design was also present in the experiment by Inman et al. (2016), it is sufficiently unusual to raise the question of whether it was responsible for the unexpected outcome of the experiment. In order to explore this possibility, the design of Experiment 2 was modified so that each group received the same amount of exposure to S+ and S–. Four groups of eight rats received training with a loud, 82 dB, and a soft, 58 dB, clicker. The stimuli were presented in an alternating sequence, and the duration of each

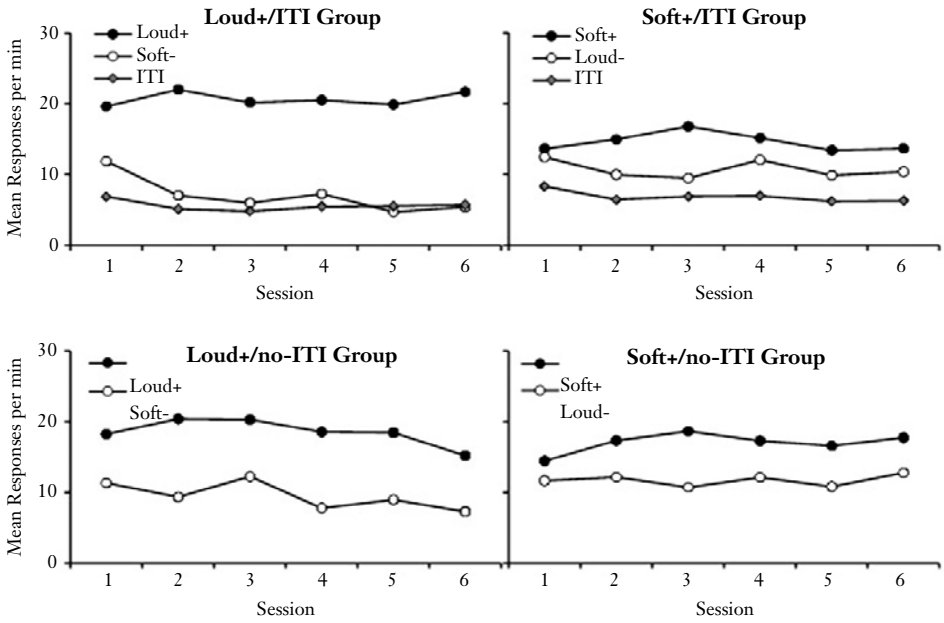

FIGURE 4. The mean rates of responding to S+ and S– for every session of Experiment 3 for two groups trained with an ITI (upper panels) and the two groups trained without an ITI (lower panels).

presentation of S+ and S– was 73 sec. On reinforced trials, food was presented according to the same schedule that was used for Experiment 2. The training for the loud+/ITI group and the soft+/ITI group was identical to that for their namesakes of Experiment 2, with an ITI of 73 sec separating each trial. The training for the loud+/no-ITI and the soft+/no-ITI groups was also based on that for their namesakes from Experiment 2, except that there was no interval between successive presentations of the loud and soft clicker, and the duration of S– was 73 sec rather than 219 sec. The absence of the ITI meant that the session duration for the groups trained with an ITI was twice that for groups trained without an ITI.

The mean rates of responding recorded during the experiment are presented in figure 4, which reveals a similar outcome to the experiment to that for Experiment 2. The acquisition of a discrimination between a loud and soft clicker progressed more readily when the loud clicker, rather than the soft clicker was paired with food. Moreover, this difference was found both with, and without an ITI. The discrimination ratios presented in figure 5, confirm

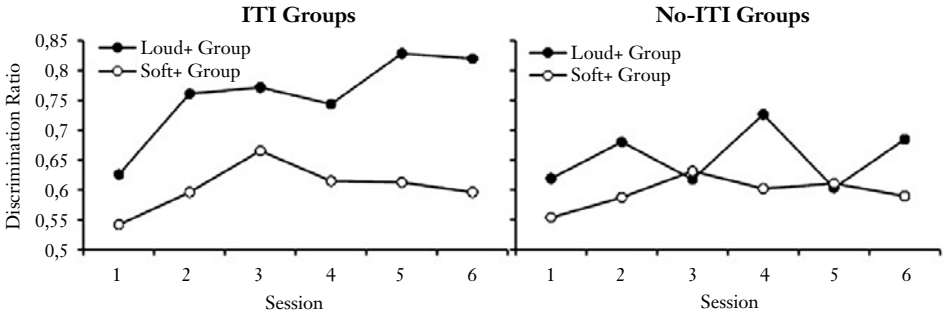

Figure 5. Discrimination ratios for each session of Experiment 3 for groups trained with an ITI (left-hand panel) and for the groups without an ITI (right-hand panel).

this asymmetry. A similar ANOVA of the ratios to that for the previous experiment revealed a significant main effect of stimulus intensity, $F(1,28) = 14.20$, $p = .001$, $\eta_p^2 = .34$. None of the interactions involving this factor was significant, ps > .10. Confirmation of an asymmetry in the No-ITI groups is provided by the significant difference between the discrimination ratios for the two groups on Session 2, $t(14) = 2.9$, $p = .011$, $r = .61$.

By demonstrating an asymmetry in a discrimination based on stimulus intensity, even when the trials are not separated by an ITI, the results from Experiments 2 and 3 strongly suggest this effect does not depend upon the generalisation of inhibition from cues present during the ITI. The purpose of the next experiment was to lend support to this conclusion, by testing it in a rather different way.

Experiment 4

Two groups of eight rats were given an intensity discrimination involving a medium intensity, 64-dB, and a soft intensity, 49-dB, clicker. The design was based on Experiment 1. The duration of the clicker was 15 sec on reinforced and non-reinforced trials, and the trials were separated by an interval with a mean duration of 6 min (range 4-8 min). Sucrose solution was presented at the end of each trial with the medium but not the soft intensity clicker for the medium+ group, and at the end of each trial with the soft but not the medium clicker for the soft+ group. In contrast to the previous experiments, the clicker that was used for the training trials was presented

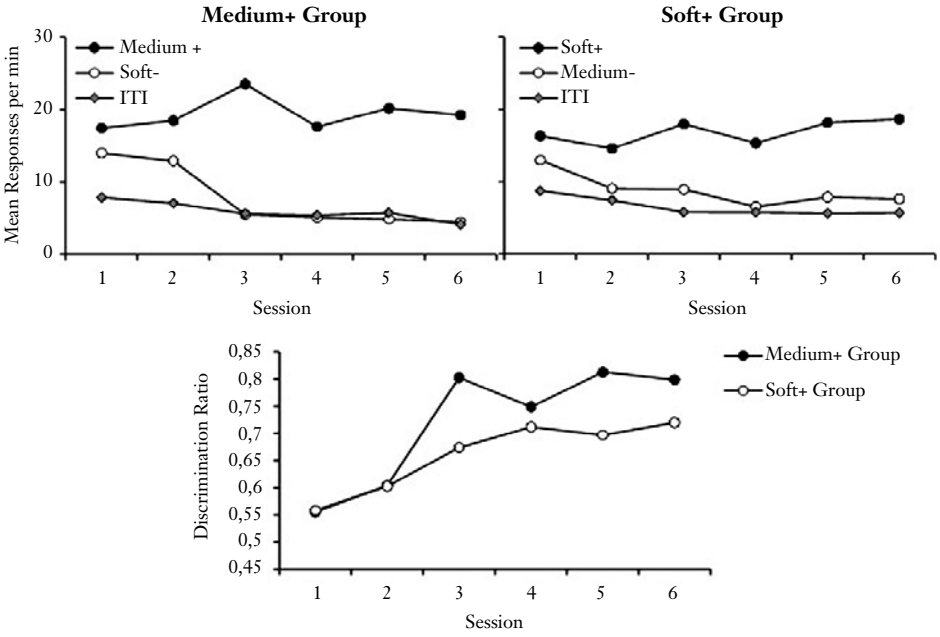

FIGURE 6. The mean rates of responding to S+ and S– for every session of Experiment 4 for the medium+ group (top left panel) and the soft+ group (top right panel) and the discrimination ratios for both groups (lower panel).

throughout the ITI, at an intensity of 82 dB.This treatment was intended to allow the loud ITI clicker to enter into an inhibitory association, which would be expected to generalise to a greater extent, by virtue of being more similar, to the medium rather than the soft intensity training clicker. As a consequence of this generalisation, the discrimination should be acquired more readily by the soft+ than the medium+ group. In other words, an asymmetry in the acquisition of the intensity discrimination would be expected, but in the opposite direction to that observed for the groups trained with an ITI in Experiments 1-3.

The results from the six sessions of discrimination training can be seen in figure 6, where it is evident from the top row that the foregoing prediction was not confirmed. Indeed, in keeping with the previous experiments, the medium+ soft– discrimination was acquired more readily than soft+ medium–. The lower panel of figure 6 depicts the same results plotted as discrimination ratios. A two-way ANOVA of individual mean discrimination ratios for each

of the six sessions of training revealed a significant main effect of group, $F(1,14) = 8.24$, $p = .012$, $\eta_p^2 = .37$. The remaining findings for this analysis were a significant main effect of session, $F(5,70) = 15.26$, $p < .001$, $\eta_p^2 = .52$, and a non-significant Session × Group interaction, $F(5,70) = 1.62$, $p > .10$.

Discussion

The results from the four experiments join a number of previous reports by showing an asymmetry in the acquisition of a discrimination based on auditory intensity. A discrimination between an intense and a weak stimulus progressed more readily when the reinforcer was signalled by the intense, rather than the weak stimulus. This effect has previously been shown in rats, either with different intensities of white noise for conditioned suppression (Jakubowska & Zielinski, 1976; Pierrel, Sherman, Blue, & Hegge, 1970; Zielinski, 1965), or with different intensities of a tone for a free-operant appetitive discrimination (Blue, 1967; Sadowsky, 1966). It has also been shown in humans with different intensities of a tone for eyelid conditioning (Moore, 1964). The present experiments extend these findings by demonstrating the asymmetry for the first time with different intensities of a clicker for appetitive Pavlovian conditioning with rats. More importantly, from a theoretical perspective, the present experiments have shown that the asymmetry occurs when there is no interval between successive training trials. The present experiments also demonstrate for the first time that the asymmetry in discriminations based on intensity remains evident when there is an ITI that is filled with a stimulus that is of greater intensity than the more intense of the two training stimuli. As noted above, these last two findings make it difficult to attribute the asymmetry to the generalisation of inhibition from cues present during the ITI. How then can the results be explained?

Stimulus intensity dynamism

An obvious starting point for a discussion concerning the present results is Hull's (1952) proposal that the strength of a conditioned response to a stimulus is determined not only by the strength of the CS-US association, but also by the intensity of the CS. These two factors were assumed to interact

in a multiplicative fashion, along with the level of Drive, to determine response strength. This influence of stimulus intensity was referred to as Stimulus Intensity Dynamism which, Hull further suggested, will result in an asymmetry in discriminations based on magnitude.

> When the simple discrimination of two stimulus intensities occurs, the difference between the intensities remaining constant, the process is more effective in terms of the net reaction potential (${}_{S}E_{R}$) yield when reinforcement is given to the more intense rather than to the less intense of the two discriminanda.[2]

At the start of this chapter, a quotation from Mackintosh (1974, pp. 532-3) was presented in which he summarised an alternative to stimulus intensity dynamism for the effects of stimulus intensity on conditioned responding (see also Perkins, 1953; Logan, 1954). The relative merits of these contrasting accounts were considered in an extremely thorough review by Gray (1965). He concluded that a clear effect of stimulus intensity on responding was not reliably found when the stimulus was presented continuously throughout the experimental session. An effect was found, however, when the stimulus was presented sporadically with individual trials separated by an ITI. This pattern of results is exactly that predicted by the proposals of Perkins (1953) and Logan (1954). Gray therefore concluded that the influence of stimulus intensity on conditioned responding is due to the processes of generalisation and discrimination, rather than to stimulus intensity dynamism. A similar conclusion was reached by Mackintosh (1974, pp. 43-44), and by Moore (1964) when explaining his demonstration of an asymmetry in a discrimination based on the intensity of a tone.

It is not possible on the basis of the four experiments described in this chapter to reject completely an explanation for our results in terms of stimulus intensity dynamism. Nonetheless, given the conclusion drawn by Gray (1965), Mackintosh (1974), and Moore (1964), together with the lack of evidence that unambiguously demonstrates this effect, it would seem prudent to seek an alternative explanation for our findings.

2. Hull, 1952, Theorem 17B, p. 87.

Feature Positive Effect and the Rescorla-Wagner (1972) theory

Kosaki, Jones and Pearce (2013) demonstrated that rats find it easier to locate one of two submerged platforms in a rectangular swimming pool, if the platforms are situated near the centres of the long walls (but not the short walls) than if the platforms are situated near the centres of the short walls (but not the long walls). They regarded this finding as a further demonstration of the asymmetry that is found with magnitude discriminations, where the two magnitudes were provided by the lengths of the walls of the pool. Thus a long+ short– discrimination was acquired more readily than short+ long–. In order to explain this outcome it was suggested that the lengths of the walls were represented as a set of elements, with a long wall containing the same elements as a short wall together with elements that are unique to the long wall. At its simplest, this explanation would stipulate that a short wall is represented by element A, and a long wall by elements A and B. Such a conceptualisation then leads to the long+ short– discrimination being characterised as AB+ A–, and short+ long- as A+ AB–.The asymmetry in the discrimination based on length can then be considered to be a demonstration of the feature-positive effect, in which a discrimination involving two stimuli is acquired more readily when the distinctive feature, B, is present on the reinforced trials, AB+ A–, than on the non-reinforced trials, A+ AB– (e.g. Hearst, 1978; Jenkins & Sainsbury, 1970). A similar explanation was put forward earlier by Todd, Winterbauer, and Bouton (2010), to explain their demonstration of an asymmetry in the acquisition of a magnitude discrimination based on stimuli of different durations.

To return to the present experiments, it is not unreasonable to extend the above analysis to discriminations based on stimulus intensity. A weak intensity clicker might be regarded as comprising a single element, A, while a louder clicker might be regarded as comprising elements A and B. The asymmetry recorded in the reported experiments can therefore be regarded as a further manifestation of the feature-positive effect.

The next step in this analysis is to explain why a feature-positive discrimination, AB+ A– should be easier to acquire than a feature-negative discrimination, A+ AB–. To address this matter, Bouton and Hendrix (2011) turned to the Rescorla-Wagner (1972) theory. According to this theory, when two or more stimuli are presented for conditioning, the opportunity is provided for

each of them to enter independently into an association with the unconditioned stimulus (US). Moreover, the strength of the response to the compound is directly related to the sum of the associative strengths of its elements. Given these principles, it follows that the rate at which each discrimination is solved will be determined by how quickly B acquires its associative properties. These properties will be excitatory for the AB+ A– discrimination and, at least at the outset of training, the theory predicts they will be acquired quite rapidly with the result that the AB+ A– discrimination will also develop rapidly. As far as the A+ AB– discrimination is concerned, by virtue of being presented only on non-reinforced trials, B will gain negative associative strength that will result in a weaker response to AB than A. However, the rate at which this negative strength is acquired is directly related to the excitatory strength of A, and since this will be weakest at the outset of training, it follows that the acquisition of inhibition by B and the acquisition of the A+ AB– discrimination will initially be slow.

At first sight, therefore, the Rescorla-Wagner (1972) theory appears well placed to explain the asymmetry that occurs with intensity discriminations. Moreover, it follows from the theory that the feature positive effect does not depend upon the trials being separated by an ITI, which allows it to explain the results from Experiments 2 and 3. Closer inspection of this explanation for our findings, however, reveals a potentially serious problem with it. Consider the simple modification to an A+ AB– discrimination of adding a common cue to both trials, AC+ ABC–. According to the Rescorla-Wagner (1972) theory, the acquisition of negative associative strength by B will now be faster than when it is employed for an A+ AB– discrimination, with the consequence that the discrimination will develop more rapidly when C is present rather than absent on both trials. This prediction may seem counterintuitive, as a manipulation that can be said to enhance the similarity of the signals for the presence and the absence of the US, by adding the same element to both of them, is anticipated to facilitate rather than hinder the discrimination between them. It should thus not be surprising to discover there is evidence to indicate the prediction is wrong. In an autoshaping experiment with pigeons, Pearce and Redhead (1993) demonstrated that an A+ AB– discrimination was acquired more readily than an AC+ ABC– discrimination, where the three stimuli were small rectangles of different colours presented on a television screen. In a rather different study, Kosaki, Jones and Pearce (2013) required rats to escape from a swimming pool, by finding a submerged

platform that was situated in front of a short rather than a long panel. When the two panels were 15 and 45 cm, then the discrimination was solved, but when these panels were both extended by 55 cm (70 vs 100 cm) then the discrimination became impossible to solve. The original discrimination can be characterised as A+ AB–, where A and B are respectively 15 cm and 30 cm of panel length; while the new discrimination can be characterised as AC+ ABC–, where A and B retain the same values and C is 55 cm. Given this characterisation, the Rescorla-Wagner (1972) theory again makes the wrong prediction about the experiment, by anticipating the opposite outcome. Moreover, it is difficult to find an alternative characterisation of the way in which the training stimuli are represented that will allow the theory to make the correct prediction. In view of these setbacks for an interpretation of our results in terms of the Rescorla-Wagner (1972) theory, we propose a rather different explanation.

The Feature Positive Effect: A Configural Analysis

When confronted with any kind of discrimination, Pearce (1987, 1994, 2002) proposed that rather than learn about the significance of individual elements, subjects learn about the significance of the configuration of stimulation that is present on each trial. Although we shall show that this theory is unable to explain our results, it is possible that a simplified version of the theory might be more successful. In order to introduce the simplification, it is necessary to say a few words about the original theory.

The network in Panel a of figure 7 shows the connections that will be formed during a trial with AB+. Presenting A and B together will result in input units for A and B being connected to a configural unit for AB which, in turn, will elicit a response that is appropriate to the outcome paired with the configuration. Learning in this part of the network will continue until each of A and B provide half of the activation that is necessary to excite fully the AB unit. If A is then presented by itself (figure 7b), it will activate the AB unit to half its maximum level and result in a conditioned response of half the strength observed in the presence of AB.

The box enclosing the input units in figure 7a merits some comment. Pearce (1994) proposed that the level of activation of an input unit, which

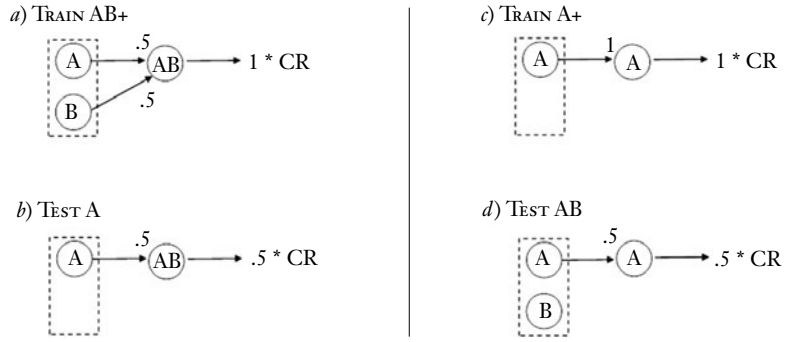

Figure 7. The connections that are predicted to be formed by the theory of Pearce (1994) after training with AB+ (Panel a), and A+ (Panel c) and the way in which these connections are predicted to be effective when a test with A follows training with AB+ (Panel b), or a test with AB follows training with A+ (Panel d).

determines its influence on the configural unit, is affected by the number of input units that are currently activated. The greater this number, the weaker will be the activation of each input unit. The presence of the box is meant to highlight this interaction between the input units. For reasons that need not concern us here, this interaction between input units does not affect the predictions that have just been made, but its significance will be made clear shortly.

Figure 7c shows the connections that will be formed during a simple A+ trial. An input unit for A will become fully connected to a configural unit for A, which will enable A to elicit a strong CR whenever it is presented. If a test trial is conducted with AB, however, then the response that is elicited will be weaker than to A (see figure 7d). The explanation for such a generalisation decrement rests with the interaction that will take place when two input units are activated simultaneously. The presence of B will restrict the activation of the input unit for A by a half and thereby restrict its capacity to elicit a response by a half.

Thus the configural theory of Pearce (1994) predicts that after conditioning with A and testing with AB, or after conditioning with AB and testing with A, the strength of the response on the test trials will be the same in both cases. In other words, generalisation between A and AB is predicted to be symmetrical and the feature-positive effect should not occur. This outcome, it is worth noting, is a consequence of the interaction that occurs when two or more input units are excited simultaneously.

When the foregoing analysis is applied to the present results, two obstacles arise. The first is that Pearce (1987, 1994, 2002) failed to provide any account for how the similarity between two intensities of the same stimulus might be determined. An obvious response to this shortcoming is to follow the line taken above, and to assume that a weak stimulus activates a small set of distinctive elements, while a strong stimulus activates a larger set that subsumes those activated by the weak stimulus. Although this proposal may be a step in the right direction, it highlights the second obstacle. If a weak stimulus is characterised as A, and a strong stimulus as AB, then the two discriminations will be A+ AB– and AB+ A–. We have just seen that the theory predicts these discriminations will be symmetrical, which was not the case for the above experiments. If the theory is to explain our results, then a different means for predicting generalisation between an intense and a weak stimulus must be sought.

One possibility is to accept the characterisation of these stimuli as elements, but to consider that manner in which a pattern of stimulation activates a configural unit is less complex than that proposed by Pearce (1994). Rather than the level of activation of each input unit being affected by the presence of concurrently activated input units, it is possible that each input unit is equally effective when it is activated either alone or in the presence of other activated inputs. To explore this possibility, figure 7 was modified by removing the boxes around the input units in order to indicate that activity in one input unit no longer has any influence on other input units (see figure 8). As figure 8a shows, this modification will mean that A and B must still compete for their connection with the AB unit during an AB+ trial. As a consequence, after training with AB, a trial with A alone will still excite a CR that is half the magnitude of that excited by AB (figure 8b). Of course, the removal of the interaction will have no influence on what occurs during an A+ trial, as shown in figure 8c. The influence of the modification will, however, be significant when, after conditioning with A+, the presence B is added to create a test trial with AB. Figure 8d shows that despite the presence of B, A will still be able to activate fully the configural unit for A, and lead to no change in the strength of the CR elicited by A.[3]

3. Thus, in contrast to Pearce (1994), it follows from the new proposals that presenting a novel stimulus with a CS will not influence the strength of the CR elicited by the latter. The many demonstrations of external inhibition (e.g. Pavlov, 1927), of course, contradict this prediction and some alternative explanation to that offered by Pearce (1994) must be found for

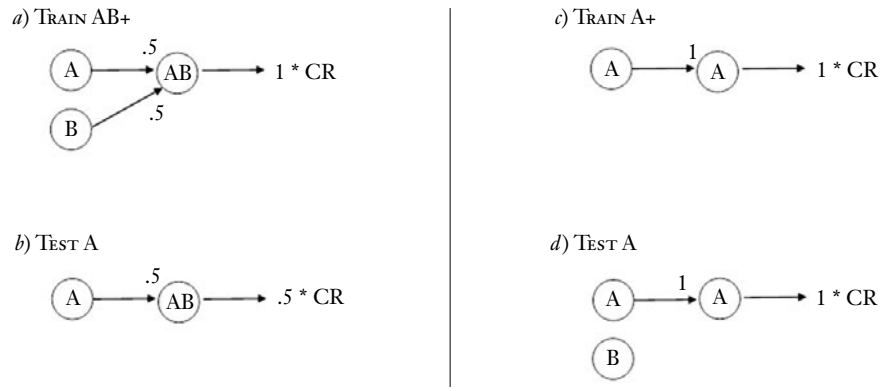

Figure 8. The connections that are predicted to be formed by the modified theory of Pearce (1994) after training with AB+ (Panel a), and A+ (Panel c), and the way in which these connections are predicted to be effective when a test with A follows training with AB+ (Panel b), or a test with AB follows training with A+ (Panel d).

The theory now predicts the feature-positive effect. The generalisation decrement that results from conditioning with AB, and testing with A, will ensure that an AB+ A– discrimination is acquired readily, whereas that lack of such a decrement after conditioning with A and testing with AB will ensure that an A+ AB– discrimination is harder to acquire.

These ideas can be represented formally. Equation 1 shows that the degree of generalisation from a training pattern to a test pattern, ${}_{train}G_{test}$, is determined by the number elements common to both patterns, N_c, expressed as a proportion of the total number of elements in the training pattern, N_{train}.

$$ {}_{train}G_{test} = N_c / N_{train} \qquad 1 $$

Following from the principles put forward in Pearce (1987, 1994), Equation 2 shows how the strength of the response to a test pattern, E_{test}, will be determined by the sum of its own associative strength, V_{test}, together with the associative strength that generalises to it from a training pattern.

$$ E_{test} = V_{test} + {}_{train}G_{test} * V_{train} \qquad 2 $$

this effect. One possibility was offered by Pavlov (1927), who attributed external inhibition to the disruptive effect of an investigatory reflex that can be elicited by novel cues.

Equation 3, then, shows how the associative strength of a stimulus, A, will change on a single trial, where β is a learning rate parameter with a value between 0 and 1, and λ is the asymptote for conditioning.

$$\Delta V_A = \beta^* (\lambda - E_A) \qquad 3$$

In order to confirm that the modified theory is able to explain the feature-positive effect, a computer simulation was conducted in which the acquisition of an AB+ A– and an A+ AB– discrimination was compared, without an ITI. The simulation was based on Equations 1-3, with β set at .2 for all trials and with λ set at 1 for reinforced trials and 0 for non-reinforced trials. The results from the simulation can be seen in the upper left-hand panel of figure 9. Although the predictions for the reinforced trials of the discriminations are superimposed, it is quite clear that the AB+ A– discrimination is predicted to be acquired more readily than A+ AB–.

We suggested above that a loud+ soft- discrimination can be characterised as AB+ A–, and a soft+ loud– discrimination as A+ AB–. If this suggestion is accepted then the foregoing simulation demonstrates that the modified configural theory is able to explain the results from Experiments 2 and 3 in which an asymmetry in the discrimination of intensity was observed without an ITI. In order to explore the predictions made by the modified theory when these discriminations involve an ITI, a further simulation was conducted. The soft+ loud– discrimination was represented as C– AC+ ABC– and loud+ soft– as C– AC– ABC+, where C represents the contextual cues that were present throughout the experiment, including the ITI. The results from the simulation are portrayed in the lower, left-hand panel of figure 9. In keeping with the results from the experiments, the loud+ soft– discrimination is predicted to be acquired more readily than soft+ loud–.

Experiment 4 revealed that a medium+ soft– discrimination was acquired more readily than medium– soft+, when a loud clicker was presented throughout every ITI. In order to determine the predictions made by the modified theory for this task, the soft clicker was represented as A, the medium clicker as AB, and the loud clicker during the ITI as ABC. The outcome of the simulation is shown in the upper right-hand panel of figure 9. During the early stages of the discrimination there is a clear advantage to the medium+ soft– task, but with extended training this outcome is predicted to reverse. While the results from the experiment confirmed the first of these predictions, it is

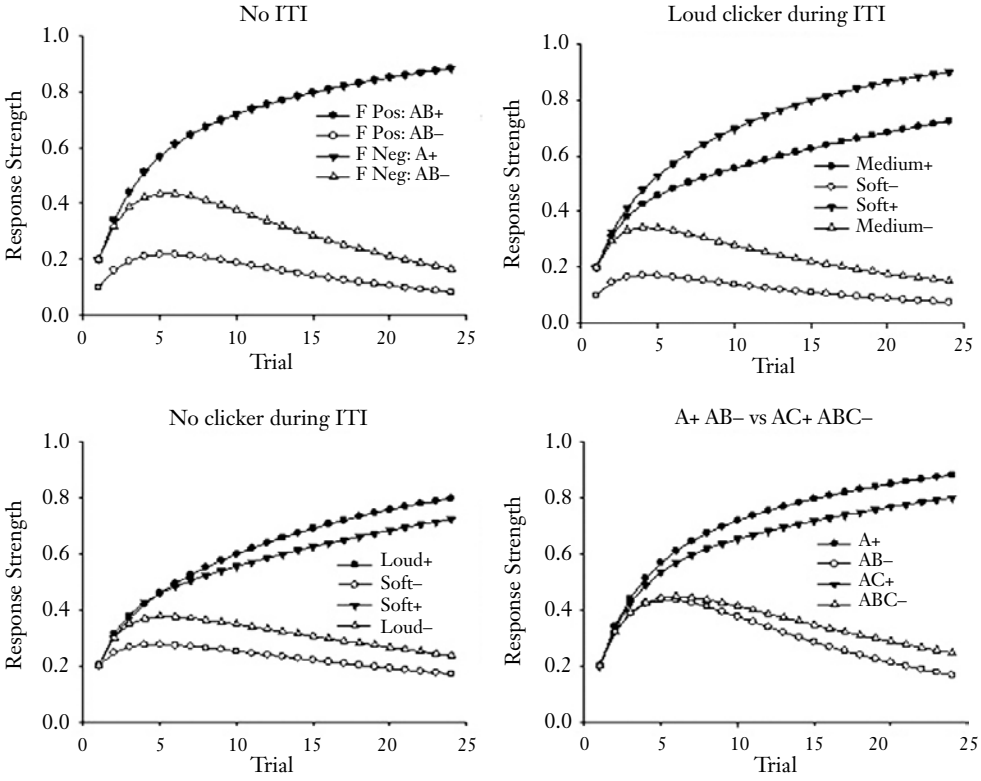

FIGURE 9. The results from computer simulations of the modified theory of Pearce (1994) for a feature positive (F Pos) and a feature negative (F Neg) discrimination without an ITI (top left panel), for a loud+ soft– discrimination and a soft+ loud– discrimination with a clicker that was silent during the ITI (bottom left panel), for medium+ soft– discrimination and a soft+ medium– discrimination with a clicker that was loud during the ITI (top right panel), and for an A+ AB– and an AC+ ABC– discrimination (bottom right panel).

unfortunate that further training was not included in the experiment to determine if, eventually, the soft+ medium– discrimination would be superior to medium+ soft–.

We have argued that a major stumbling block for an analysis of the present results in terms of the Rescorla-Wagner (1972) theory is its incorrect prediction that an AC+ ABC– discrimination will be acquired more readily than A+ AB–. A further computer simulation was conducted in order to determine the predictions made by the modified configural theory concerning these discriminations. The results can be seen in the lower right-hand panel of

figure 9. In keeping with experimental findings (e.g. Pearce & Redhead, 1993; Kosaki et al., 2013) the theory predicts that the A+ AB– discrimination will be acquired more readily than AC+ ABC–.

It would seem, therefore, that the simplification to the theory of Pearce (1994) allows it to explain quite well the results described above. The present chapter is not the place to explore further the implications of this modification. However, apart from the problem posed by external inhibition alluded to above, a preliminary exploration of the predictions it makes has, so far, failed to reveal a major conflict with existing experimental findings.

Implications for the Discrimination of Quantity

The present experiments were conducted in order to examine if discriminations based on stimulus intensity are solved in the same manner as discriminations based on differences in quantity. In fact, the results have shown that these discriminations are solved in rather different ways. A possible reason for this outcome is that quantity is represented in a different manner to intensity. Thus, while it seems likely that an increase in intensity results in the addition of elements to those already recruited by the weaker stimulus, an increase in quantity may have a different effect. Perhaps the dimension of quantity is similar to more conventional dimensions such as the wavelength of light, or frequency of sound. Movement along these dimensions might result in the activation of previously inactive elements, as well as the inactivation of previously active elements, with the result that the number of elements activated by any particular value on the dimension remains relatively constant (e.g. Blough, 1975). According to this analysis, any two stimuli on the same dimension will each activate a set of shared elements, and a set of unique elements, so that they can be treated as AC and BC. All three of the accounts considered above predict that in the absence of an ITI, a discrimination between these patterns will progress at the same rate irrespective of which one signals the reinforcer. Thus, if two different quantities are also characterised as AC and BC, the foregoing analysis can readily explain why an asymmetry with a quantity discrimination does not occur in the absence of an ITI (Inman et al., 2015; Inman et al., 2016).

To explain the asymmetry that is seen when the stimuli for a quantity discrimination are separated by an ITI, we might assume that the dimension is anchored with the value of zero, and that this value will be activated during the ITI. Thus zero might be represented by ABC, a small quantity by BCD, and a large quantity by CDE. A small+ large– discrimination can then be regarded as ABC– BCD+ CDE–, whereas the opposite discrimination can be regarded as ABC– BCD– CDE+. For both tasks, the non-reinforced exposure to ABC will disrupt excitatory conditioning with BCD to a greater extent than CDE and result in a large- small+ discrimination being acquired more slowly than large+ small–. Both the amended configural theory and the Recorla-Wagner (1972) theory make this prediction.

A troubling aspect of these proposals is that little has been said about the nature of the features that are excited when a particular quantity is presented. In the case of a discrimination where quantity is indicated by different numbers of black squares on a white screen (e.g. Inman et al., 2015), it is tempting to suggest that elements are related to the amount of black, or the amount of white, on the display screen. However, Inman et al. (2015) found that a quantity discrimination based on black squares against a white background was barely affected when the stimuli were changed to white squares on a black background. This finding suggests the elements that represent numbers do not relate to a concrete property of the training stimuli, such as the amount of black or white, but to a more abstract property. At present, it is hard to specify what this abstract property might be.

In conclusion, by showing that the asymmetry in discriminations based on stimulus intensity is not abolished either by removing the ITI, or by manipulating the stimuli present during the ITI, the present experiments point for the first time to the possibility that the asymmetry observed in magnitude discriminations occurs for two reasons. We have suggested that these reasons may stem from the different effects that changes in magnitude can have on the elements activated by a particular stimulus. An increase in the intensity of an auditory cue may simply add to the elements that are activated by the original cue, whereas an increase in quantity might not affect the number of elements that are activated, but might affect the type of elements that are activated.

References

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 104*, 3-21.

Bouton, M. E., & García-Gutiérrez, A. (2006). Intertrial interval as a contextual stimulus. *Behavioural Processes, 71*, 307-317.

Bouton, M. E., & Hendrix, M. C. (2011). Intertrial interval as a contextual stimulus: Further analysis of a novel asymmetry in temporal discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 79-93.

Gray, J. A. (1965). Stimulus intensity dynamism. *Psychological Bulletin, 63*, 180-196.

Hearst, E. (1978). Stimulus relationships and feature selection in learning and behaviour. In S. H. Hulse, H. Fowler, & W. K. Honig (Eds.), *Cognitive processes in animal behaviour*. Hillsdale, N. J.: Erlbaum.

Hull, C. L. (1952). *A behaviour system*. New Haven: Yale University Press.

Inman, R. A., Honey, R. C., & Pearce, J. M. (2015). Asymmetry in the discrimination of magnitude: The role of stimulus generalisation. *Journal of Experimental Psychology: Animal Learning and Cognition, 41*, 309-321.

Inman, R. A., Honey, R. C., Eccles, G. L., & Pearce, J. M. (2016). Asymmetry in the discrimination of quantity by rats: The role of the intertrial interval. *Learning & Behavior, 44*, 67-77.

Jakubowska, E., & Zielinski, K. (1976). Differentiation learning as a function of stimulus intensity and previous experience with the CS+. *Acta Neurobiologiae Experimentalis, 36*, 427-446.

Jenkins, H. M., & Sainsbury, R. S. (1970). Discrimination learning with the distinctive feature on positive or negative trials. In D. Mostovsky (Ed.), *Attention: Contemporary theory and analysis* (pp. 239-273). New York: Appleton-Century-Crofts.

Kosaki, Y., Jones, P. M., & Pearce, J. M. (2013). Asymmetry in the discrimination of length during spatial learning. *Journal of Experimental Psychology: Animal Behavior Processes, 39*, 342-356.

Kyd, R. J., Pearce, J. M., Haselgrove, M., Amin, E., & Aggleton, J. P. (2007). The effects of hippocampal system lesions on a novel temporal discrimination task for rats. *Behavioural Brain Research, 187*, 159-171.

Logan, F. A. (1954). A note on stimulus intensity dynamism (V). *Psychological Review, 61*, 77-80.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London: Academic Press.

Moore, J. W. (1964). Differential eyelid conditioning as a function of the frequency and intensity of auditory CSs. *Journal of Experimental Psychology, 68*, 250-259.

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

Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review*, *94*, 61-73.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587-607.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior*, *30*, 73-95.

Pearce, J. M., & Redhead, E. S. (1993). The influence of an irrelevant stimulus on two discriminations. *Journal of Experimental Psychology: Animal Behavior Processes*, *19*, 180-190.

Pelz, C., Gerber, B., & Menzel, R. (1997). Odorant intensity as a determinant for olfactory conditioning in honeybees: Roles in discrimination, overshadowing and memory consolidation. *The Journal of Experimental Biology*, *200*, 837-847.

Perkins, C. C., Jr. (1953). The relation between conditioned stimulus intensity and response strength. *Journal of Experimental Psychology*, *46*, 225-231.

Pierrel, J., Gilmour Sheman, J., Blue, S., & Hegge, F. W. (1970). Auditory discrimination: A three-variable analysis of intensity effects. *Journal of the Experimental Analysis of Behavior*, *13*, 17-35.

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.

Todd, T. P., Winterbauer, N. E., & Bouton, M. E. (2010). Interstimulus interval as a discriminative cue: Evidence of the generality of a novel asymmetry in temporal discrimination learning. *Behavioural Processes*, *84*, 412-420.

Vonk, J., & Beran, M. J. (2012). Bears "count" too: Quantity estimation and comparison in black bears (Ursus Americanus). *Animal Behaviour*, *84*, 231-238.

Watanabe, S. (1998). Discrimination of "four" and "two" by pigeons. *Psychological Record*, *48*, 383-391.

*Nicholas J. Mackintosh and the Renaissance of Animal Psychology in Spain: A Collaborative Enterprise**

Gabriel Ruiz

Department of Experimental Psychology,
Universidad de Sevilla, Spain

Abstract. The aim of this chapter is to explore the scientific importance and personal significance of Nicholas Mackintosh in the origin and further development of the Spanish Society for Comparative Psychology. In order to achieve this, I will take into account the evolution of Spanish Psychology during the 1970s and 1980s from a philosophical, theoretical and local discipline, mainly concerned with applied problems, to a more international and sophisticated discipline committed with scientific research. Therefore, my contribution will have three different parts. Firstly, I will recall the first steps of animal psychology in Spain during the 1920s. Secondly, I will propose some reason why these former and promising developments disappeared during the Spanish Civil War (1936-1939) and during Franco's times (1940-1975). Finally, I will move from the mid-70s to the late 80s, a time in which a new generation of young psychologists interested in animal learning and comparative psychology arose — Luis Aguado, Gumersinda Alonso, Santiago Benjumea, Victoria Díez Chamizo, Francisco Fernández Serra, Víctor García-Hoz Rosales, Matías López, Antonio Maldonado, Helena Matute and Ricardo Pellón. Through canvassing the thoughts and experiences of those Spanish psychologists who spent time in the labo-

* Some fragments of this chapter were presented in *Tribute to N. J. Mackintosh* given at the XXVII International Congress of the SEPC (Universidad de Sevilla, 9-11 September 2015). I am profoundly grateful to Natividad Sánchez in particular for her help and comments without which this work would not have been possible. I am also grateful for the help of Luis Aguado, Gumersinda Alonso, Santiago Benjumea, Isabel de Brugada, Victoria Díez Chamizo, Francisco Fernández Serra, Víctor García-Hoz, Felisa González, Antonio Guillamón, Matías López, Francisco J. López, Antonio Maldonado, Helena Matute, Ricardo Pellón and Juan Manuel Rosas. I have taken seriously everybody's advice, but the interpretative decisions are my own, and I of course bear full responsibility for them and for any errors that remain. Author's email: gruiz@us.es.

ratories of Sussex and Cambridge, Nicholas Mackintosh appears as one on the major influences on this new generation that was looking for a new scientific approach to the study of mind and behaviour, far away from the old-fashioned philosophical and theoretical point of views supported by the Spanish psychologists during the 60s and 70s.

Introduction

For a historian of psychology whose career started as a researcher in the field of animal learning and cognition, I feel privileged to be able to contribute to this book paying tribute to Nicholas J. Mackintosh. My contribution does not aim to be a history of this field of psychology in Spain, or a review of the experimental and ethological studies into animal learning and behaviour that have been conducted here in recent years. Instead my intention in this work is to explore the personal significance and the scientific impact that Mackintosh has had in the renaissance of animal psychology in Spain and I will try to show how far he influenced the first generation of Spanish researchers in animal learning and cognition who started studying in the 1980s. This generation has played a decisive role in the institutionalization of animal psychology in Spain, starting up many of the research laboratories that are working today and founding the *Sociedad Española de Psicología Comparada* [Spanish Society for Comparative Psychology — SEPC in its Spanish acronym], both of which have played a pivotal role in the revitalization of this research area in Spain over the last twenty years.

To do this, I have asked many of the people who formed part of this generation to tell me about the experiences they had with Mackintosh and I have used that information to construct a historical account. The story starts long before the founding of the SEPC and Mackintosh appears as a key figure in the formation of the way we think and research today. Seeing him in this light, as part of our own history, is also my tribute to his memory.

The first experiments on animal learning in Spain (1900-1936)

The nineteenth century was not a good period for science in Spain. After a long cycle of wars and colonial setbacks, the Spanish people, impoverished

and mostly illiterate, were under the yoke of a Catholic church which controlled political power and education, and they were governed by an aristocracy more interested in maintaining their privileges than in modernising the country. Universities were in the hands of the state and their intellectual elites lived with their backs to European philosophical and scientific thinking. The imperviousness of this intellectual backwater gradually gave way, not without much resistance, and positivism, materialism, Darwinism and experimental sciences began to make inroads in Spain during the few periods of peace and freedom which the country enjoyed during those years (Sánchez Ron, 1999).

However, the first third of the twentieth century was a period of vigorous intellectual and scientific activity in Spain known as the *Silver Age of Spanish Letters and Sciences*. This "silver age" got underway in 1906 when Santiago Ramón y Cajal (1852-1934) received the Nobel Prize for Physiology/Medicine, and came to a traumatic end in the summer of 1936 with the outbreak of the Civil War, when General Francisco Franco (1892-1975) staged an uprising with the support of part of the army against the republican government. There had been a growth in international contacts during three decades of progress which had laid the institutional foundations which would help modernize Spanish science. Particularly important was the creation in 1907 of the *Junta para la Ampliación de Estudios e Investigaciones Científicas* [Board for the Advancement of Studies and Scientific Research — JAE in its Spanish acronym]. The most prominent Spanish scientist at the time, Santiago Ramón y Cajal, was its president from 1907 until 1934, and during his presidency the JAE completed two important tasks: 1) it set up a grant system enabling students, lecturers and researchers to work abroad; and 2) it promoted the creation of research centres and laboratories throughout the country so that the researchers with grants could teach and research when they came back to Spain.

Although the JAE's grant policy helped promote scientific psychology in Spain, its impact on the development of studies into animal learning was very limited. From the total of almost 2,000 grants awarded up to 1938, when the institution disappeared, 6.1% had requested the grant to study questions related to psychology. However, most of these researchers were interested in clinical or educational subjects, so they chose centres in countries such as France, Switzerland and Germany, where animal psychology was not the field of greatest interest (Mateos & Blanco, 1997).

As for the second of the JAE's objectives, the creation of large scientific institutions, several research centres were set up, amongst which we should

highlight the following: one dedicated to humanities, another to physical-natural sciences, which included Cajal's own *Laboratorio de Investigaciones Biológicas* [Laboratory of Biological Research] and, finally, a group of physiological institutes located in the *Residencia de Estudiantes* [Students' Residence]. The most notable of these were the *Instituto de Fisiología General* [Institute of General Physiology] of Juan Negrín (1892-1956), the *Instituto de Fisiología y Anatomía de los Centros Nerviosos* [Institute of Physiology and Anatomy of the Nerve Centres] of Gonzalo Rodríguez Lafora (1886-1971) and the *Instituto de Histología Normal y Patológica* [Institute of Normal and Pathological Histology] of Pío del Río-Hortega (1882-1945). One can see that none of these laboratories were dedicated specifically to psychology. This absence of psychological research laboratories was not exclusive to the JAE because the presence of psychology in Spanish universities was negligible at the start of the twentieth century.

The first professorship of Psychology which existed at the Universidad de Madrid was that of Philosophical Psychology, created in 1898 in the Faculty of Philosophy and Letters. The teaching of psychology was included in the degrees of Philosophy, Medicine and Sciences. Soon afterwards, in 1902, the professorship of Experimental Psychology was created in the Faculty of Sciences at the Universidad de Madrid, where the teaching of psychology was included in the doctorate in Medicine and Philosophy. The professorship of Philosophical Psychology was always kept on a scholastic philosophical perspective, far from the advances in experimental psychology. However, Luis Simarro (1851-1921), the first professor in Experimental Psychology at a Spanish university, did have a major influence on the development of scientific psychology in Spain, particularly through some of his pupils such as Gonzalo Rodríguez Lafora, the first Spaniard to publish experimental studies on animal learning (Moya, 1986).

Rodríguez Lafora, a neurologist interested in the problem of brain localizations, was the first Spaniard to publish experiments which used conditioning techniques. Lafora finished his studies in medicine in 1907 in Madrid, having acquired a solid background in the anatomy of the nervous system with Simarro and in his collaboration at Ramón y Cajal's *Laboratorio de Investigaciones Biológicas*. Thanks to a JAE grant, he gained more experience in Berlin (1908) with Oskar Vogt (1870-1959) and Theodor Ziehen (1862-1950), and in Munich (1909) with Emil Kraepelin (1856-1926) and Alois Alzheimer (1864-1915). In 1910, he was appointed neuropathologist at *Saint Elizabeth's Government Hospital for the Insane*, in Washington D.C., and there he worked

with Shepherd Ivory Franz (1874-1933) in his studies about the localisation of cerebral functions. The most important of these collaborative works was *On the Functions of the Cerebrum: The Occipital Lobes* (1911). This was an experimental study on the cerebral bases of perception, in which monkeys learnt several discriminations and then they evaluated the level which minor or major ablations of the occipital lobes produced associative visual impairments. Rodríguez Lafora became familiar with the conditioning techniques used by Franz, amongst them the puzzle-boxes, and after his return in 1912, he set up the first research programme in Spain using instrumental conditioning procedures.

Between 1915 and 1936, Rodríguez Lafora and his disciples performed studies on the experimental physiology of the nervous system, such as the one they performed on the function of the corpus callosum in monkeys and cats, in which they used puzzle-boxes to determine the effects of these injuries on the instrumental motor behaviour of these animals. To do this they conditioned the animals so they learnt to execute different movements with each hand (monkeys) or forepaw (cats) and they observed the effects that the lesion of the corpus callosum had on this learning (see figure 1a and 1b).

The main result of these experiments was that the lesion of the corpus callosum produced paretic and "apraxic" phenomena of the opposite side of the body, which disappeared completely after about twenty days. From this moment on, the animals were once again able to perform the movements they had learnt before the operation. They also saw two more important results: 1) the existence of a direct relation between the extension and importance of crossed symptoms and the extension and depth of the lesion of the corpus callosum; 2) the involvement of a remote action of inhibition, also produced by the lesion of the corpus callosum on the nearest motor centres which, in turn, produced the "apraxic" and motor symptoms on the opposite side of the body (Rodríguez Lafora & Prados, 1921).

In that same year, 1921, Santiago Ramón y Cajal published the first work of comparative psychology conducted in Spain, a study about the sensorial capacities of ants. After reviewing the works of the main authors who had investigated these questions (e.g. Lubbock, Fabre, Forel, André, Bethe, Piéron, Cornetz, Bouvier, etc.), Cajal conducted multiple experiments in which he demonstrated that these hymenopterans did not discriminate colours, they only discriminated the different brightness of one single colour, black. Their experiments showed that the olfactory and tactile perception of ants was,

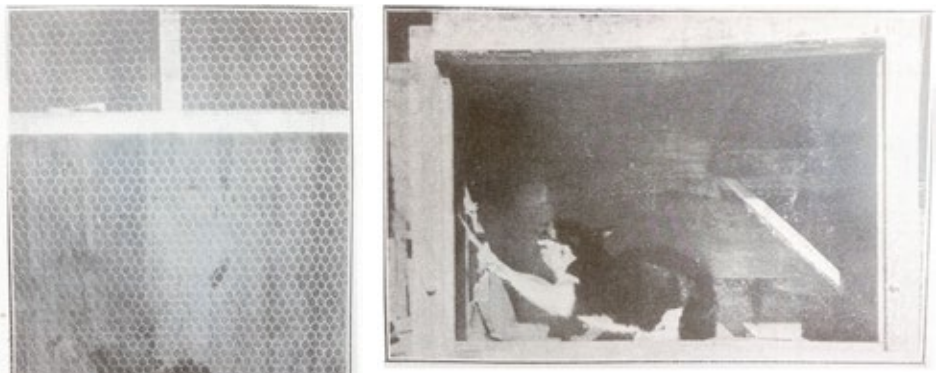

FIGURE 1. Photos of the puzzle-boxes used by Rodríguez Lafora in his studies into the function of the corpus callosum. In both cases, monkeys (a) and cats (b) performed an action with their left hand or foot (pulling on a ring tied to a string), and they had to keep the string tight to be able to take the food with their right hand or foot. To make these boxes, Rodríguez Lafora drew inspiration from those used by Berry (1908) and Haggerty (1909) in their studies into imitation in cats and monkeys, respectively.

however, excellent. Cajal concluded his study stating that the brain of ants was a prodigiously complex and subtle associative machine which compensated for their poor sensorial world:

> In conclusion: ants...endure great sensorial hardship. Apart from sense of touch and smell, which is highly developed in them, the other senses give the animals confusing and fragmentary observations of the outside world...
>
> And yet, these insects compensate for this sensorial impoverishment with a prodigiously rich variety of motor reactions and the most marvellous instincts. The senses are not the most important aspect of psychic life: above them, coordinating the information and interpreting it in the light of age-long acquisitions of the species, dominates the brain, so rich in its potential.[1]

1. Ramón y Cajal, 1921, p. 571.

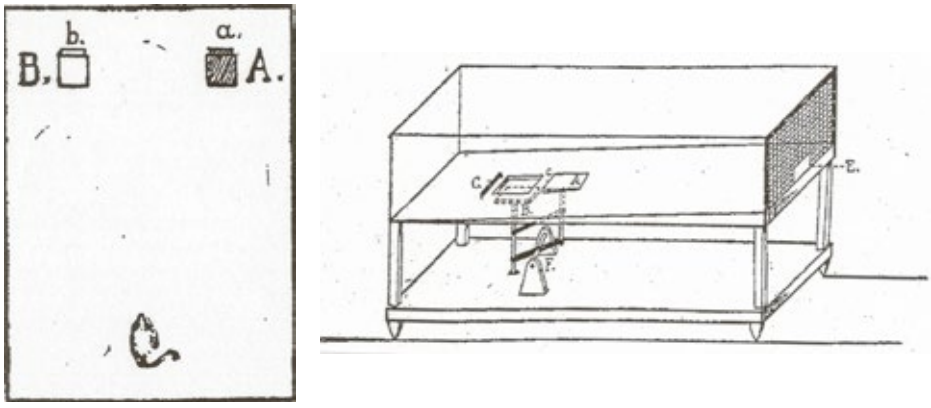

Figure 2. Experimental situations used by de Luna in his studies: a) discrimination experiments: the mouse had to discriminate between the goal box containing the food (A) and the one without food (B); the number of goal boxes, their colour, the distance between them and the distance between the starting point and the goal boxes varied in different experiments; the access to the goal boxes was from the opposite side (a or b) to where the animal was at the beginning of each trial, to prevent the food from being seen or smelt from that position. b) Puzzle-box designed by de Luna for the study of associative memory in small rodents, he described how it worked as follows: "On the slightly sloping floor of the cage whose entrance is at E, there are two holes, where platforms A and B of rocker arm F appear. C is a little curtain which, when extended, is held weakly in the hook c of plate A. The weight of the mouse which is on this last plate makes it drop slightly (until it hits a stop which cannot be seen in the figure), just enough for the curtain to roll back through the action of a spring and the food appears on B" (de Luna, 1921, p. 396).

Also in 1921, Joaquín de Luna, one of Cajal's students, published a study about discriminative learning in mice, shortly after finishing his medical studies in Madrid (de Luna, 1921). He used the following task: the mouse had to discriminate the goal box (upper case A or B), which contained the food and enter the box on the opposite side (lower case a or b) to where it was at the start of the trial (see figure 2a). The colour and position of the goal boxes varied and de Luna studied the differences in learning depending on the type of mouse (albino, grey and mixed) and its sex, and the distance between and the number of goal boxes (two, four or six). He also wrote about a puzzle-box which he had designed to study associative memory: "to manage to eat, the mouse, previously deprived of food, must stand on platform A, which drops a little [until a spring is released in C] and the food appears on platform B" (de Luna, 1921, pp. 396-397) (see figure 2b).

Unfortunately, these studies did not continue because, after finishing his MD in 1922, De Luna moved to Paris in the following year to work at the *Laboratory of Comparative Embryology* under the direction of Louis-Félix Henneguy (1850-1928) at the Còllege de France and he ended up living there permanently (Bandrés & Llavona, 1997).

As we have seen, the experiments of Gonzalo Rodríguez Lafora and Joaquín de Luna used procedures of instrumental conditioning. When did Pavlovian research get underway in Spain?

At the start of the twentieth century, Ivan P. Pavlov (1849-1936) presented his theory of conditioned reflexes to the world at the XIV International Congress of Medicine held in Madrid in 1903. Although interest in Pavlovian reflexology had been present in Spain since the end of the nineteenth century in the work of the physiologists José Gómez Ocaña (1860-1919), Ramón Turró i Darder (1854-1926) and August Pi i Sunyer (1879-1965), the first Spanish Pavlovian is considered to be the military doctor Galo Fernández-España (1854-1933). He wrote a series of articles on the reflexology of Pavlov and Vladimir M. Bechterev (1857-1927), which were published in the *Revista de Sanidad Militar* [Journal of Military Health] between 1914 and 1924 (Bandrés & Llavona, 1996).

In 1929, the second Russian edition of the book *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* was translated into Spanish. Pavlov had published it in 1927, and the Spanish translation included a prologue dedicated to Spanish readers where the Russian physiologist expressed his satisfaction that the public of this language could have access to "a topic of general intellectual interest...and even, from a certain point of view, of vital practical interest" (Pavlov, 1929, p. VII). The translation of this volume had been promoted by the Spanish doctor Gregorio Marañón (1887-1960) and it was one of his disciples, the endocrinologist Juan Planelles (1900-1972), who published the first work in Spain using the technique of conditioned reflexes (Planelles & Luwisch, 1935).

Planelles had studied medicine in Madrid and he was already a member of the Royal Academy of Medicine by the time he was 25. With a grant from the JAE, he travelled to Germany and Holland where he learnt surgical techniques for physiological research, such as "Pavlov's pouch" and biliary fistula. On his return to Spain he was appointed professor of Therapeutics at the Universidad de Salamanca and later he concentrated on pharmacological research at the *Instituto de Investigaciones Clínicas* [Institute of Clinical Research]

in Madrid, of which he was the founder and director until the outbreak of the Civil War in 1936 (Marco-Igual, 2011; Martínez, 2014).

In his studies on the metabolism of carbohydrates, Planelles performed experiments in which he took repeated blood samples in the experimental room before feeding his dogs. This meant that the dogs associated the taking of blood samples with the arrival of the food, and these blood extractions caused a hypoglycaemic reaction to the sight of food. The dogs in the control group were never fed in the experimental room after the blood extraction and they were seen to have a hyperglycaemic reaction. This is possibly one of the first demonstrations of the contextual control of appetite (Planelles & Luwisch, 1935).

Taken as a whole, the experiments of Rodríguez Lafora on the cerebral bases of instrumental motor behaviour in monkeys and cats, Ramón y Cajal's into the sensorial functions of ants, Joaquín de Luna's into discriminative learning in mice and Planelles' into the Pavlovian conditioning of the hypoglycaemic response in dogs, are indicative of the existence of an incipient interest for the experimental study of animal learning in Spain during the first third of the twentieth century. This interest went no further for several reasons: firstly, there were no animal psychology laboratories and the only laboratory of experimental psychology that there was at the time, the one founded by Simarro in his professorship at the Universidad de Madrid, was more concerned with the dissemination of Wundtian physiological psychology and its practical applications than in the creation of new scientific knowledge (Bandrés, Llavona & Campos, 1996). Secondly, both Rodríguez Lafora and de Luna formed part of Ramón y Cajal's histological school, they were a group of researchers concentrating primarily on the histological analysis of the nervous system, and they considered purely behavioural studies as complementary or accessorial to anatomical ones. Finally, the Spanish Civil War (1936-1939) forced many of the authors we have mentioned, such as Rodríguez Lafora and Planelles, to emigrate as they were in danger of being tried and sentenced for their political ideas.

A catholic science (1939-1975)

The Spanish Civil War (1936-1939) ripped the country apart at all levels and Franco's new authoritarian regime was built on a National-Catholic vision of

the world. To start with, the scientific institutions created during the Republic were dismantled, as occurred, for example, with the JAE which had been directed by Ramón y Cajal. The JAE was replaced by the *Consejo Superior de Investigaciones Científicas* [Spanish National Research Council — CSIC in its Spanish acronym], a new public research body which the Franco government set up to foster a "Spanish science". The first president of the CSIC was José Ibáñez Martín (1896-1969), who was also National Education minister in Franco's government from 1939 to 1951. In his opening speech, he left no doubt as to his understanding of the expression "Spanish science":

> Science is for us an aspiration towards God. We want a Catholic science. Let us, therefore, at this time, be rid of all those scientific heresies which have dried out... the channels of our national genius and sunk us into apathy and decadence...Our current science, in connection with what in previous centuries defined us as a nation and as an empire, wants to be above all Catholic.[2]

Important decisions were taken to reform education. The first was to instigate a purge of teachers who were considered responsible for having inoculated the Marxist virus in society and in young minds (Morente, 2001). This purge was aimed at eliminating teachers who had sympathised with leftist ideas and select in their place teachers of absolute moral and Catholic solvency, who would be loyal to the new Franco state. Every aspect of the public, professional and private lives of teachers and lecturers was investigated and a large number of them were tried and sentenced, which could mean a fine, disqualification from work, prison or worse. Many of the prosecution files included charges such as "belonging to the Freemasonry", "being a layman", "[having demonstrated] that they did not have the strength that an educator of the young heroes of Spain must have", "writing scientific works to give the foreigner a feeling of normality in the area under the control of the Marxist government", "boast publicly about being a Darwinist" (Otero, 2006). Rodríguez Lafora, for example, was accused of being of a "markedly leftist ideology" and sentenced to eight years of disqualification from work and the payment of a heavy fine; in 1938 he went into exile in Mexico.[3] The

2. Martínez-Arias, 2009, p. 1183.

3. During his time in Mexico, Rodríguez Lafora carried on researching and publishing, but about clinical subjects far removed from the experimental studies we have described in

situation of Planelles was more serious because he had been a member of the Communist Party of Spain and Undersecretary of Public Health in the Republican government, thus his life would have been in danger had he been arrested, and in 1937 he went into exile in the USSR.[4] Many of Ramón y Cajal's disciples were also tried and disqualified from their work; they had to give up their research and work as doctors to survive; others chose exile (Otero, op. cit.).

After the purge process had got underway, the next step was educational reform: the subordination of teaching and the dissemination of science to the Catholic religion; this was particularly damaging for Darwinism. The subject of evolution, which was on the baccalaureate syllabus before the war, was removed, to be replaced by a creationist interpretation of biology which accepted, literally, the story of Genesis (Blázquez, 2011). In one official text from *Ciencias Naturales de segundo de bachillerato* [Second Year Baccalaureate Natural Science], it was stated that God created the four kingdoms of nature separated by abysses rendering any evolutionist explanation impossible (Muedra, 1955) (see figure 3).

In another book, *A Dios por la ciencia. Estudios científico-apologéticos* [To God through science. Scientific-apologetic studies], the author wanted Baccalaureate students to see:

> God in the works of Nature which are the works of his hands and raise your heart, with the help of the marvels you see around you, to his hidden and invisible ones...To know an artist, a man of science, one has to study his works...To know God one has to see him in his works.[5]

In the same book, there was a chapter with the heading *¿Tienen inteligencia los insectos?* [Do insects have intelligence?] After describing the social organization of bees and their skill in constructing the hive, the author, a Jesuit priest, concluded:

---

this article. He returned to Spain in 1947, but did not regain his previous posts until his purification sentence had run its course in 1949. He retired in 1955 (Lafuente, Carpintero & Ferrándiz, 1991).

4. Unlike Rodrígez Lafora, Planelles could not fulfill his wish to return to Spain and he died in the USSR in 1972, after playing an important role in pharmacological research.

5. Simón, 1947, p. 10.

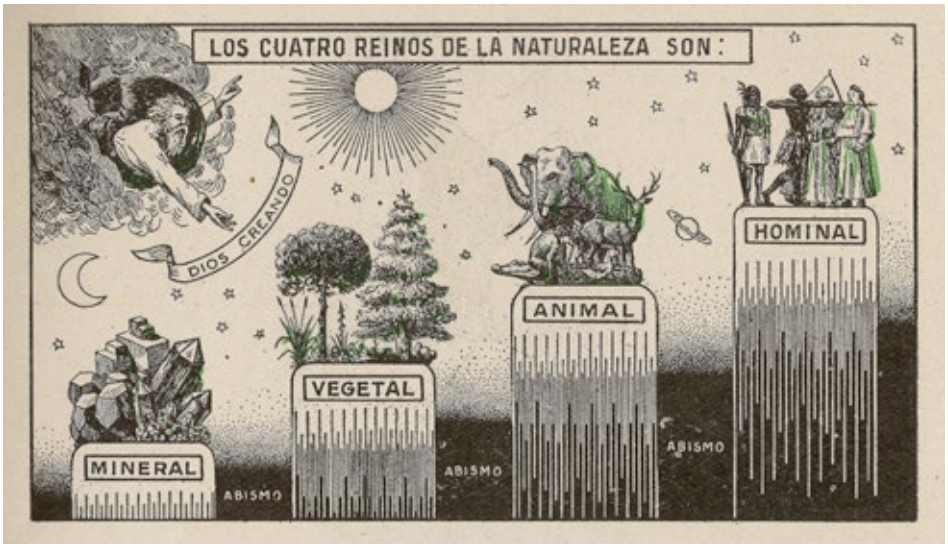

FIGURE 3. Figure included in an official baccalaureate text published in 1955 showing God creating the four kingdoms of nature: mineral, vegetal, animal, human. The author, Vicente Muedra, a Jesuit priest, stated: "We can imagine all natural beings situated on four steps or stairs, one on top of the other, but without being able to pass from one to the other" (Muedra, 1955, p. 12). Regarding man, Muedra states the following: "Finally, on the top step, occupying the highest point on these stairs, is man, the king of creation, dominating all other beings. It is impossible for those on the lower steps, in other words, animals to get to where man is, because the abyss which separates them, is impenetrable and no bridge or connection can bring them closer; and definitely not unite them. Man has an immortal and intelligent soul which the animal neither has nor will have: this is the abyss which separates both groups and which will always keep the distances between them. For this reason, man has intelligence. In contrast, the animal does not have intelligence, which is why it does not think or progress" (Muedra, 1955, p. 13).

> Either bees are talented mathematicians, experienced at resolving each day problems that would be beyond the reach of the vast majority of men; or...they are prodigious chemists who know how to manufacture the cleverest of substances; or...they are consummate architects, with a perfect knowledge of the laws of statics; or...they have brilliant minds for social and state affairs, geniuses of economics and forecasts...or one has to recognise the intelligence of one who is above them and to whose impulse they move, [in the same way] that the hand of the child, who does not know how to write, moves and writes under the impulse of his father...Which of these two extremes do we accept?...You can see that the first is inadmissible, the bee is one of the little beasts which, beyond the honey-

> comb and its constructions, looks like one of the most stupid. It does not even bear comparison with the fly...No; bees do not have understanding...There is another mind, another intelligence which directs them and of which they are no more than blind executors. A wise mind which knows...mathematics, which knows the laws and reactions of chemistry, which knows the laws of statics. That mind is the mind of God.[6]

One of the authors who did most to divulge this pre-Darwinian natural theology was Vicente Muedra, biologist, Jesuit priest and a prolific author of books on animal behaviour, such as the one entitled *La perfección científica en las obras animales. Narraciones científico-recreativas* [Scientific perfection in animal works. Scientific-recreational narratives] (Muedra, 1948), in which he stated: "We would be extremely satisfied if...we had achieved closer approximation of the reader to God, helping him to see the Creator in the admirable works of his creatures" (pp. 8-9). The argument defended by Muedra was that animals resolved a multitude of problems in an instinctive way and these instincts were the manifestation of the wisdom and science of the superior being which had created them. In another of his books, *Maravillas científicas en los actos animales* [Scientific marvels in animal acts], he took the feeding habits of the shrike to warn young Spaniards about the cruelty of the Bolsheviks (see figure 4):

> The shrike is one of the birds of prey with the cruellest and bloodiest instincts. In days gone by executioners enjoyed destroying the bodies of martyrs...the Bolsheviks and their secret police have gone one step further: they don't kill, no, they torment their victims slowly to make their agony even more unbearable. The shrike fixes on thorns...of bushes, those animals which are its favourite delicacy. Its poor victims go through hours and hours impaled on spikes, not yet dead, in an agony which is as horrible as it is prolonged. These birds are the Bolsheviks par excellence![7]

This Catholic vision of science affected experimental psychology which also disappeared from the baccalaureate syllabus. On 14 April 1939 the Official State Bulletin published the new law for the reform of secondary education and one of its articles read as follows:

6. Simón, 1947, pp. 222-223.
7. Muedra, 1950, p. 31.

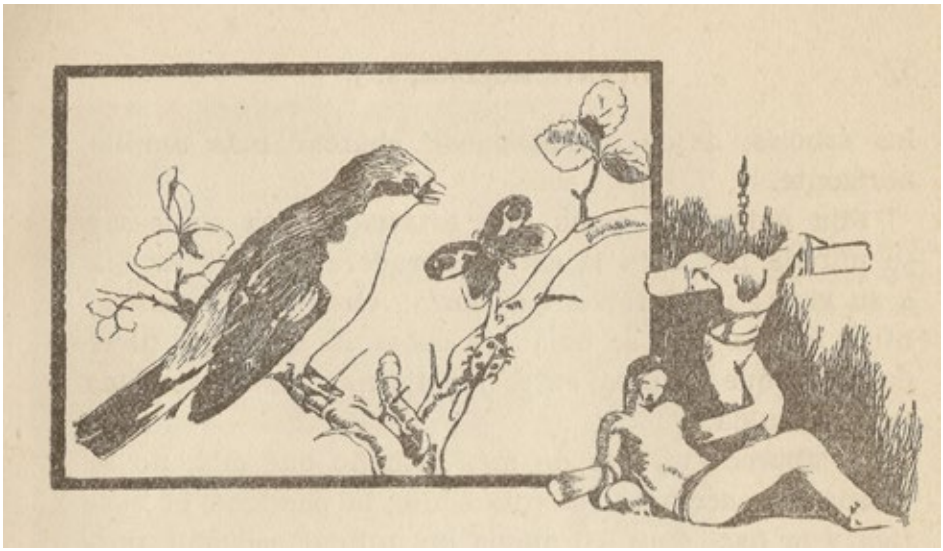

FIGURE 4. Figure which appeared in the book *Maravillas científicas en los actos animales* [Scientific marvels in the acts of animals], published by the Jesuit priest Vicente Muedra in 1950. In it he compared the feeding habits of the shrike with the torment suffered by Catholic "martyrs" in Spanish prisons controlled by communists (known as *checas*) during the Spanish Civil War. Muedra stated that shrikes: "are the "Bolshevik" birds par excellence: their mission is to kill, assassinate...but not from one blow, but only after terrible suffering. Tormenting the prey which is to constitute its food!" (Muedra, 1950, p. 30).

> As will be seen by reading the subjects of psychology, the problems referring to "experimental psychology" do not occupy a special place. It has been decided that the practice of laboratory experiments in psychology, while having value for the empirical development of this science, tends on the other hand to disorientate secondary school students, and instead of providing them with the bases for ulterior speculations, distracts them with facts that they do not know how to interpret correctly, thereby diverting them from what is most important and educational.[8]

In a context such as this, it is hardly surprising that the most influential individuals in Spanish psychology during those years were religious men such as the Dominican Manuel Barbado (1884-1945) and not scientists such as Simarro, Rodríguez Lafora or Planelles. Father Barbado assumed all the roles

8. Tortosa, Civera & Esteban, 1998, p. 546.

of academic and research responsibility, for example the professorship of Experimental Psychology of the Universidad de Madrid and the Chair of the Institute of Philosophy "Luis Vives" of the CSIC. He also wrote an *Introducción a la psicología experimental* [Introduction to experimental psychology] (1943) in which the two most quoted authors were Saint Thomas (191 times) and Saint Albert the Great (68 times). In this book he stated the following:

> Badly shall the thinking of the nation be governed, insofar as philosophical questions are concerned, if each teacher has his own doctrinal system...As we are dealing with Spain...the philosophical doctrine which must be taught in the official professorship is that contained in traditional philosophy...which is the only one accepted by the church and the only one which may provide the basis for a solid religious culture.[9]

Despite this oppressive and suffocating atmosphere, Spanish psychology began to make its way in higher education. In 1943 the *Department of Experimental Psychology* was created in the CSIC. This department formed part of the Institute of Philosophy of Father Barbado and it was directed by José Germain (1897-1986), a neurologist who had studied under Rodríguez Lafora. The scientific activity conducted in the department was particularly eclectic on a theoretical level and its orientation was fundamentally applied: they assessed a multitude of psychological tests, and collaborated with the army and many Spanish companies in selection procedures for qualified personnel (Huertas, Padilla & Montes, 1997).

Germain gathered round him a group of young psychologists who would become, at the start of the 1960s, the first professors of psychology in the Spanish university system after the Civil War: Mariano Yela (1921-1994), José Luis Pinillos (1919-2013), Miguel Siguán (1918-2010) and Jesusa Pertejo (1920-2007), amongst others. Thanks to the important work of this group, the first Schools of Psychology were set up in Madrid (1953) and Barcelona (1964). These were post-graduate schools which taught graduates of all subjects and provided them with a diploma which enabled them to exercise professionally as psychologists. Finally, in 1968, the Sections of Psychology were set up in the universities of Madrid and Barcelona.[10]

9. Carpintero, 1994, pp. 266-267.

10. This process of institutionalization culminated in 1979 with the creation of current Faculties of Psychology and the *Colegio Oficial de Psicólogos* [Spanish Psychological Association].

In line with the eclectic and markedly applied nature of Spanish psychology in those years, the psychology which was taught in the universities had a strong humanistic and psychotechnical component. The presence of experimental psychology in students' education was virtually symbolic, hardly any notions of experimental design were given and there were no laboratories. Furthermore, because of the political regime university students had to take what was known as the "three Marys": religion, politics and physical education. This is how it is remembered by Pío Tudela, professor of Experimental Psychology at Universidad de Granada and one of the first students of psychology at the Universidad Complutense de Madrid:

> The psychology we were taught was dominated by a differential perspective in terms of method...Psychological tests represented the dominant technology and the knowledge of their structure, reliability and validity gave theoretical meaning to the completion of two years of elementary statistics. The experiment as a research instrument was mentioned in some subjects and its contribution to illustrated psychological research in a specific subject of experimental psychology, but we did not receive specific teaching on experimental design nor did we have laboratories in which we could start to experiment.[11]

Given such a bleak outlook for the training of researchers, some younger teachers started to search abroad for this new wave of experimental psychology which was not fostered in their home country at all. Hence, for example, Ramón Bayés played an important role in the dissemination of the work of B. F. Skinner from Barcelona. He built one of the first artisan Skinner boxes to be used in Spain and his pigeons, *Orlando* and *Griselda*, were the first Spanish doves to "suffer" reinforcement schedules and resolve complex tasks such as Matching to Sample (Bayés, 1972, 1974, 1975).

Bayés had collaborated with the *Laboratori de Conducta* [Behaviour Laboratory], a laboratory which had been set up in 1973 under the direction of Lluís García Sevilla and Adolf Tobeña of the Universidad Autónoma de Barcelona.[12] This laboratory was located firstly in the Hospital de Sant Pau and

11. Tudela, 2010, p. 77.

12. However, the first animal laboratory in Spain was the *Laboratorio de Conducta Operante* [Operant Behaviour Laboratory] set up by Pere Julià in 1970, and which was at the Faculty of Philosophy and Letters at the Universidad Autónoma de Barcelona. Also during those years, J. M. Costa Molinari had created a *Laboratorio de Psicología Experimental* [Experimental Psychol-

then moved to the Universidad Autónoma de Barcelona, where it became an important centre in the training of the first generations of animal psychologists in Catalonia. The *Laboratori* disappeared in 1980; the subjects researched there included different areas such as the verification of Eysenck's theory, the influence of aversive stimuli on behaviour, the obtaining of psycho-physiological measures of bilingualism, studies on the reliability of measurements of electrical resistance of the skin, operant conditioning and its applications in different fields, the regulator variables of avoidance behaviour, and intracranial electrical stimulation (Bayés & Garau, 1982).

Although the political atmosphere was not favourable for anything coming from Russia, in the mid-1960s the psychiatrist Antonio Colodrón brought the ideas of Ivan P. Pavlov back to the clinical and university circles of Madrid, unleashing the anger and reproach of Franco's academic authorities. This is how he remembers it:

> For those of you who did not live those years, you cannot imagine the feeling of hiding and secrecy that pervaded our lives. The sanctums of psychiatric National-Catholicism defined what was tolerable with such constraints that psychoanalysis itself suffered after the speech of Pope Pius XII at the International Congress of Histopathology of the Nervous System and at the Congress of Psychotherapy and Clinical Psychology the following year...As for Pavlov, having been negated for so long, when at the end of the 1950s his name was once again in circulation, a conspiracy of silence emerged. Naming him provoked a catchphrase: he is completely overrated; a relic of the past. Man is more than a dog; he is free and can say "no".[13]

At the end of the 1960s and early 1970s, Víctor García-Hoz Rosales played a decisive role in the dissemination of Clark Hull's theory of learning in his classes at the Universidad de Madrid. Pío Tudela recalls:

> [The influence of Hull] came to us through Eysenck's theory of personality which was brilliantly taught to us by Víctor García-Hoz...In that class we started to read the original articles of Eysenck and Spence and we became familiar with Hull's

---

ogy Laboratory] at the Faculty of Medicine of the same university. Both labs were working until 1972 and when they closed, they effectively merged to become the *Laboratori de Conducta*. Julià left Spain in 1973: "being unable to tolerate the university atmosphere of those years, I returned to the USA and then I went to Mexico" (see Ruiz, Pellón & García, 2006, p. 86).

13. Colodrón, 2003, p. 293.

> theory of behaviour. For some of us this meant a definitive commitment to experimental psychology and a starting point which was going to lead in a relatively gradual manner to cognitive positions.[14]

Since then, García-Hoz has been a key player in the renaissance of research into animal learning and cognition in Spain. At the start of the 1980s, with Luis Aguado and Javier Campos, he helped found the *Laboratorio de Psicología Animal* [Animal Psychology Laboratory] at the Universidad Complutense de Madrid, and he was its first director. As you can see in figure 5, a significant proportion of the generation of researchers who set up the SEPC and provided the driving force for this scientific area in Spain did their PhDs under his supervision.

However, the Animal Psychology Laboratory directed by García-Hoz was not the first animal psychology centre founded in Madrid. A few years earlier, at the start of 1976, Antonio Guillamón had started the *Laboratorio de Psicobiología* [Psychobiology Laboratory], which was initially located in the Department of Morphology and Neuroscience of the Faculty of Medicine of the Universidad Autónoma de Madrid.

After completing his MD, Guillamón spent time during 1973-1974 with Jeffrey Gray at Oxford and he was interested in the effect of frustrative nonreward. During the autumn of 1974, Mackintosh was invited to give a seminar at the Department of Experimental Psychology of Oxford University. His book *The Psychology of Animal Learning* had been published recently and Guillamón recalls that the chapter which Mackintosh had dedicated to extinction had helped him greatly in the design of his experiments with the runway. At the end of the seminar, he talked to Mackintosh about the effect of extinction after partial reinforcement and he gave him a signed copy of the book. After he returned to Spain, this book passed through the hands of the first students who went to his laboratory to do their degree projects or doctoral theses at the end of the 1970s, including some prominent researchers of the generation responsible for creating the SEPC, such as Victoria Díez Chamizo and Ricardo Pellón.[15]

14. Tudela, 2010, p. 68.

15. The Psychobiology Laboratory of Guillamón has also played an essential role in the development of psychobiological research in Spain. The following have studied at this lab: Santiago Segovia, Andrés Parra, Emilio Ambrosio, Mari Cruz Rodríguez del Cerro, Azucena Valencia, José María Calés and César Venero, amongst others.

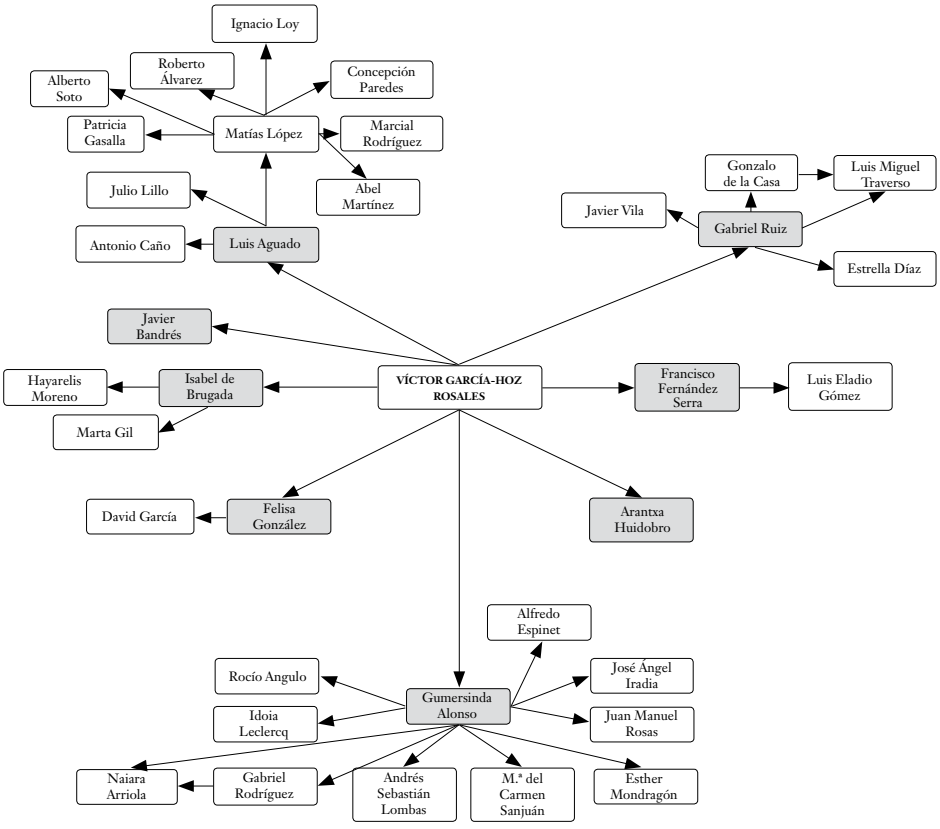

Figure 5. The influence of Víctor García-Hoz Rosales on some members of the generation which set up the SEPC, and who conducted their PhDs under his supervision (see shaded boxes). The figure only includes researchers who have worked in animal learning and cognition.

The first half of the 1970s was still a very hard time on a political and social level. The last firing-squad executions performed by Franco's regime took place on 27 September 1975, only a few months before the death of the dictator, on 20 November of that year. From the mid-1960s, workers' and students' movements had become the focal points for resistance to the regime and they felt the full force of its repression. I don't think it would be an exaggeration to say that the first years of psychology graduates learned about Pavlovian conditioning and stimulus generalization amongst strikes and assemblies demanding amnesty, freedom, democracy and the closure of Amer-

ican military bases in Spain, and running in front of the "*grises*", or greys, as we called the regime's police because of the colour of their uniforms.

The renaissance of animal psychology in Spain (1975-1985)

With the dictator's death in 1975, a process of transition commenced taking the country from the old Franco regime to a democratic system. Political parties and unions were legalized, and the autonomous regions were created. At the start of the 1980s, countercultures began to emerge, such as the *movida madrileña*, in a clear move away from Franco's society they projected an image of a "more modern" Spain, far removed from the image that the country had had for the previous forty years of dictatorship. However, large sections of the social and economic structures of the country were still Franco loyalists, and one section of the army, horrified by the legalization of the Communist Party, attempted a coup d'état on 23 February 1981 erupting into and taking over the Parliament building. Fortunately, the coup did not prosper and the democratic process continued its course.

In between songs by Radio Futura,[16] films by Almodóvar and ETA terrorism, new animal psychology laboratories opened their doors during the 1980s in Madrid (Universidad Complutense), Granada (Universidad de Granada), Sevilla (Universidad de Sevilla) and San Sebastián (Universidad del País Vasco). Many of those who would be responsible for the resurgence of animal psychology in Spain studied and worked in these labs and the ones we have mentioned already: Luis Aguado, Gumersinda Alonso, Javier Bandrés, Javier Campos, Santiago Benjumea, Victoria Díez Chamizo, Francisco Fernández Serra, Matías López, Antonio Maldonado and Ricardo Pellón.

This process of institutionalization continued in the second half of 1980s, when new labs were set up under the direction of the following researchers: Victoria Díez Chamizo at the Universidad of Barcelona (1987), Matías López at the Universidad de Oviedo (1988), Ricardo Pellón at the Universidad Nacional de Educación a Distancia in Madrid (1990), Julián Almaraz at the Uni-

16. *Radio Futura* was one of the most popular pop rock bands in Spain during the 1980s and early 1990s.

versidad de Málaga (1990), Helena Matute at the Universidad de Deusto (1991), Fernando Sánchez-Santed, Inmaculada Cubero and Pilar Flores at the Universidad de Almería (1993), Ángeles Agüero at the Universidad de Jaén (1995), Carmen Torres at the Universidad de Jaén (1999) and Juan Manuel Rosas at the Universidad de Jaén (2004) (see figure 6).

Although each laboratory developed its own research programme, the members of this generation all had something in common, something that broke with the isolation of the academic world during the Franco years: their desire to continue their training in prestigious international laboratories. An essential role was played by the different international funding programmes such as the Fullbright scholarships with the USA, British Council grants with the UK and the different post-doctoral grants and grants for overseas studies which were created in the Spanish autonomous regions. Thus, Víctor García-Hoz Rosales visited King's College Institute of Psychiatry and Sussex University, Victoria Díez Chamizo the universities of Birmingham and Cambridge, Ricardo Pellón visited Cardiff University and Antonio Maldonado, Gumersinda Alonso, Matías López and myself visited Cambridge University. These visits and many other subsequent ones helped create a network of international contacts which have made animal psychology one of the most active and internationally oriented areas of Spanish psychology.

Road to SEPC: A collaborative enterprise (1985-1989)

This network of contacts soon started to give its fruits. In March 1985 most of this group presented works at the congress *Current Perspectives in Cognitive Psychology* which was held in Madrid. At this congress, the youngest of us had the chance to discuss our research with Robert Rescorla, one of the guest speakers at the congress. In April of the same year, Victoria Díez Chamizo organized the first of a series of advanced courses on subjects of animal learning at the *Instituto de Ciencias de la Educación* [Institute of Education Sciences — ICE in its Spanish acronym] at the Universidad de Barcelona entitled *Comparative Psychology of Discrimination Learning and Modern Conditioning Theory*. This first course was given by Nicholas Mackintosh and John Pearce and was followed in successive years by two more courses: *New Perspectives in Animal Learning* given by Nicholas Mackintosh in June 1987, and *Associationism and Perceptual Learning* given by Geoffrey Hall and Nicholas Mackintosh in April 1988.

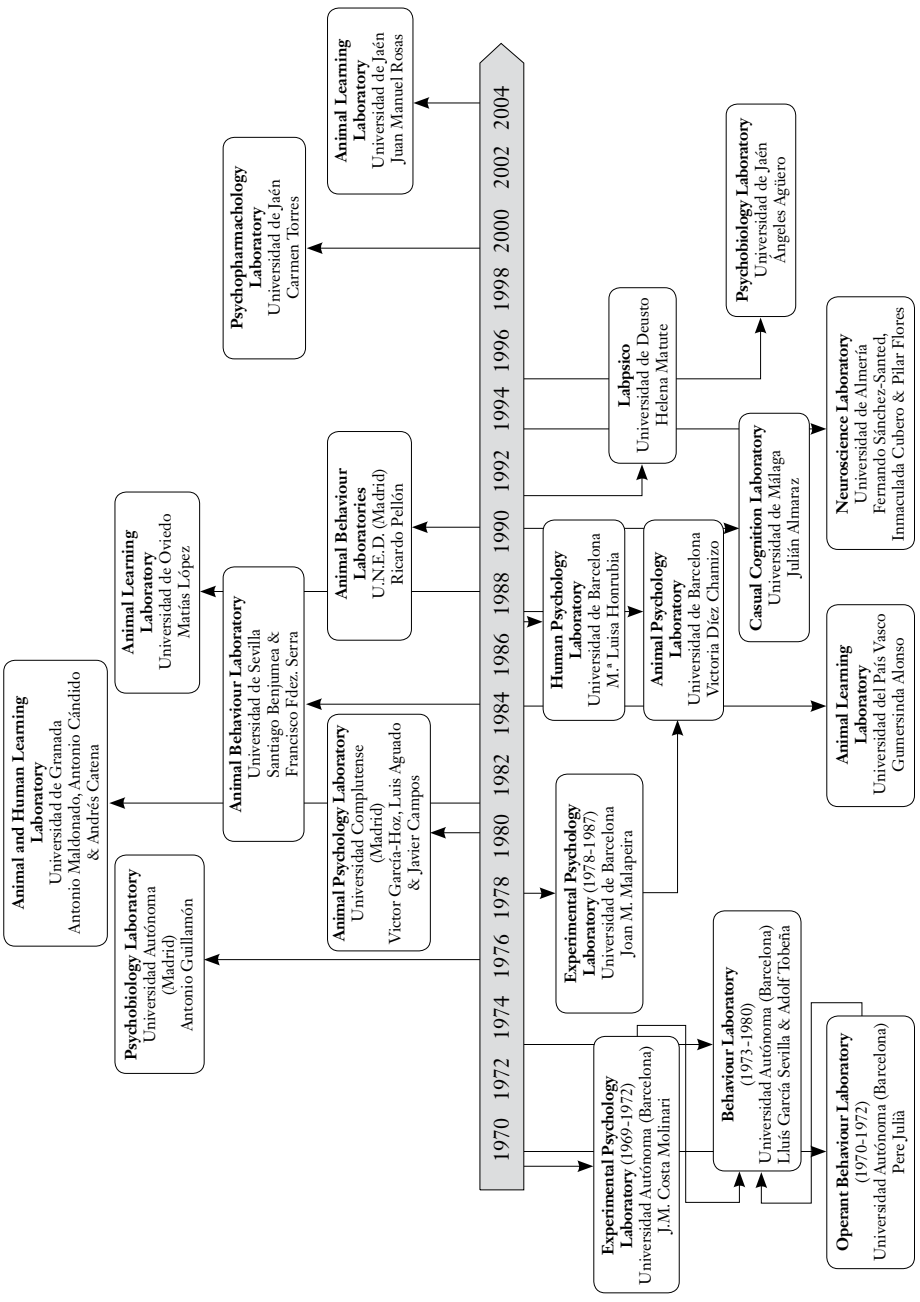

FIGURE 6. Chronology of the founding of the Spanish laboratories dedicated to animal and human learning research which have played a prominent role in the development of this research area in Spain over the last thirty-five years.

These courses strengthened even further the bonds between Spanish and British colleagues and many of us there began to get the feeling of a group, in other words, we got the impression that a community of interests was beginning to take shape around research into animal learning. In this setting there was a fairly broad consensus that the associationist vision of conditioning that Mackintosh had developed in his books *The Psychology of Animal Learning* (1974) and *Conditioning and Associative Learning* (1983) was a good theoretical framework.[17] All that was needed was a forum for the exchange of research and experiences.

On 24 June 1988, scarcely one month after the last of these courses, the idea that had been circulating was put into action and Javier Campos and Javier Bandrés organized the first scientific meeting of the *Grupo Español de Psicología Comparada* [Spanish Group for Comparative Psychology] at the Faculty of Psychology of Madrid's Universidad Complutense. Eighteen researchers from most Spanish universities attended, amongst them all the group of people who had been taking part in the previous courses, and the guest speaker was, once again, Nicholas Mackintosh. At this meeting a society was constituted with the name of *Sociedad Española de Psicología Comparada* [Spanish Society for Comparative Psychology]. Nicholas Mackintosh was named as the president of honour and it was agreed that the person responsible for organising the annual congress should be the society's president (and thus it remained until Víctor García-Hoz Rosales was proposed as president in 1993 and it was decided that the organizer of the congress should act as secretary).[18]

One year later, in April 1989, the process culminated in the celebration of the congress of Granada, under the current label SEPC. Antonio Maldonado was the president of this congress and, as a result, the first president of the SEPC. Once again, Nicholas Mackintosh was one of the invited speakers, accompanied on this occasion by Anthony Dickinson. Since then, annual meetings have been held at different Spanish universities, and the growth of the society during the 1990s was accompanied by the development and consolidation of the different laboratories thanks to concerted public funding of scientific activity.

17. The Spanish version of the book *Conditioning and Associative Learning* was published in 1988, with a translation by Victoria Díez Chamizo of the Universidad de Barcelona.

18. When García-Hoz left the presidency of the SEPC it reverted to the initial system where the organizer of the annual congress is also the president of the society for that year.

FIGURE 7. Visit of Nicholas Mackintosh to Madrid in July 1987. From left to right: Gumersinda Alonso, Victoria Díez Chamizo, Víctor García-Hoz Rosales, María Luisa Velasco, Javier Bandrés, Luis Aguado, Nicholas Mackintosh, Javier Campos and Antonio Maldonado (courtesy of Luis Aguado).

CONCLUSION: NICHOLAS MACKINTOSH, AN IMPORTANT FIGURE IN OUR RECENT HISTORY

As we have seen, animal psychology in Spain got off to a slow start during the first third of the twentieth century with the studies of Rodríguez Lafora, Ramón y Cajal, de Luna, and Planelles. After the Spanish Civil War (1936-1939), a purge took place of teaching institutions which virtually ended science in Spain: many scientists and intellectuals were tried for their political ideas and excluded from their professorships and laboratories, others had to flee into exile. From 1939 to 1975, the Franco dictatorship isolated the country from any external influence and promoted a Catholic view of science in which there was no room for evolutionist ideas or experimental psychology, which were considered harmful for Spanish youth because they were materialist. After Franco's death in 1975, Spain began to open up

to modernity and a process of democratic transition commenced in which political parties and trade unions were legalised, the Constitution was approved and the territorial organisation of the state was reformed. It was during this period of freedom that the first animal psychology laboratories in Spain began their work with the first generation of Spanish researchers in animal cognition and learning such as Luis Aguado, Gumersinda Alonso, Santiago Benjumea, Victoria Díez Chamizo, Francisco Fernández Serra, Víctor García-Hoz, Antonio Guillamón, Matías López, Antonio Maldonado, Helena Matute and Ricardo Pellón.

The vast majority of those who formed part of this first generation spent time at Sussex and Cambridge University and came into contact with Mackintosh in his frequent visits to Spain. And they recognise the enormous influence of his work:

> [*The Psychology of Animal Learning*] has had a decisive influence on the conception that many of us have of the psychology of learning and on what we have taught. That is definitely my case and I believe the same goes for many others. For me it really was a bible.[19]

> During the summers [I spent in Cambridge in the mid-1980s] the first ICE courses were organised [in Barcelona in 1985, 1987 and 1988]. Nick was delighted to go (so were John Pearce and Geoffrey Hall) despite the somewhat "tight" budget...But they were so enthusiastic and so wanted to help. I remember that after their addresses they had a queue of people waiting to ask them things...And they loved it! Clarifying doubts, making suggestions, designing experiments...It was idyllic, in a way. They infected us all with their enthusiasm and energy. In my opinion they (but particularly Nick) were responsible for many of us deciding to pursue our interests and even get excited about certain subjects related to associative learning. No-one has ever influenced me as much as Nick. He set a magnificent example (and so did Anthony Dickinson), with a professional stature and generosity that would be difficult to improve on. Before that, two people had had a great influence on my professional life: José Antonio I. Carrobles and Antonio Guillamón...but no-one had as much influence on me as Nick.[20]

19. Víctor García-Hoz Rosales, Universidad Complutense de Madrid.
20. Victoria Díez Chamizo, Universidad de Barcelona.

For all of us who started to work in the area of learning with animals [in the early 1980s], [*The Psychology of Animal Learning*] was *the* book...I am sure that many of our theses stemmed directly or indirectly from the study of that work.[21]

His modern Pavlovian vision of learning, his attentional theory of conditioning (1975)...his profound associative analyses, which are reflected in his two most important books...and his comparative analysis of animal intelligence, I think, have had a great influence...on the work of all of us in recent years.[22]

My first meditated reading of a book on learning...was precisely Nick's text *The Psychology of Animal Learning*, and then...*Conditioning and Associative Learning* had a great impact on me...two works which have undoubtedly conditioned my scientific interests.[23]

At [the SEPC Congress in 1988 in Granada] there were some very important people who were publishing articles, including international journals. They included Chamizo, and Aguado, Maldonado, Pellón, Alonso, Ruiz, López, and García-Hoz. I had read things by them and they were all there. And so were Dickinson and Mackintosh! And they accepted me in the group...I was light years behind all of them, but they urged me on...That Congress, that group, with Nick, Tony and Víctor at the head, changed my life. They were the motivation and also the proof that it was possible. I saw that others had achieved it. Victoria was publishing with Nick. They were the main model. I remember above all how kind Nick was in those early years, offering to discuss things with me which must have seemed completely trivial to him, with mistakes all over the place. He even seemed to understand my version of English. I decided to give it a go.[24]

...I can say without a shadow of doubt that my interest in psychological research took root when at the end of the 1980s my lecturer of...Psychology of Learning, our dear friend Julián Almaraz, "forced" us to read what would end up being an absolutely stimulating revelation, Nick's manual, *Conditioning and Associative Learning* translated by Victoria in 1988...It was an absolute pleasure to read and was certainly responsible for making me follow a research career in psychology. Up until that moment, I wasn't all that keen on the subject to the extent that I was

21. Luis Aguado, Universidad Complutense de Madrid.
22. Gumersinda Alonso, Universidad del País Vasco.
23. Matías López, Universidad de Oviedo.
24. Helena Matute, Universidad de Deusto.

on the point of dropping out to study medicine. That it was possible to do rigorous science in the area of behavioural sciences was made clear to me by Nick in his book. It was a real pleasure and an intellectual delight which still has me hooked.[25]

No less important and lasting than his scientific influence has been his personal significance, the role that Mackintosh played in this generation as an academic model, one which was very unlike the Spanish professors of that time:

...on those visits we realised that being an important professor and an internationally renowned researcher did not mean one had to be inaccessible or disregard the ideas of a beginner. Absolutely, things were very different to what we were used to at home! I remember from the start the pints we had at the pub and the long nights at Nick's house; this opened our eyes to academic practices which were unknown [in Spain].[26]

...he went to all the weekly laboratory seminars and showed the same interest towards the presentations regardless of who was giving them, and I saw this as proof of his straightforwardness and quality as a person.[27]

...for me, Nick and Tony have been "my masters" (along with Ramón Bayés and Víctor García-Hoz Rosales) and those who have helped me most in my work, especially early on, [I have tried] to emulate them and be like them as much as possible, as I consider them without a doubt the best scientists and people I have met in my life as a professional in Psychology.[28]

At a personal level I would say that starting to work with Nick at Cambridge marked a turning point in my life. My first stay at the Department of Experimental Psychology [in Cambridge] was in 1982...The experience was highly stimulating and enriching. I was amazed by everything: the theoretical questions, experimental rigour, camaraderie and the warmth with which I was received (well, we all were). Seeing Nick so accessible, so "normal", was incredible for me.[29]

25. Francisco J. López, Universidad de Málaga.
26. Luis Aguado, Universidad Complutense de Madrid.
27. Matías López, Universidad de Oviedo.
28. Antonio Maldonado, Universidad de Granada.
29. Victoria Díez Chamizo, Universidad de Barcelona.

> Nick was one of the nicest people in psychology that I have ever met...He was, of course, a highly intelligent man, but also warm-hearted, attentive and natural, who took you to his level with never a trace of condescension. He was elegant and had a great sense of humour, very English, like his elegance, and his open, frank laughter, one of the most cheerful laughs I have ever heard. He was always ready to make a joke, sometimes at your expense, but it never felt bad.[30]

When I look around me, thirty years after the first visits of Nicholas Mackintosh to Spain, and I listen to the studies being presented at the congresses of the SEPC and I read the articles published by Spanish authors in the specialist journals, I can see clearly that the prevailing orientation in the study of learning in this country continues to be associative. And I believe that this orientation, which owes much to the ideas of Mackintosh, due to his conceptual rigour and strictly experimental nature, has been determinant for the overall development of Spanish experimental psychology.

References

Bandrés, J., & Llavona, R. (1996). Pavlov, Bechterev y el objetivismo: La psicología como la veía Galo Fernández-España, *Revista de Historia de la Psicología*, *17*, 44-53.

Bandrés, J., & Llavona, R. (1997). Joaquín de Luna y Juan Planelles: el aprendizaje y los orígenes de la psicología experimental en España. *Revista de Historia de la Psicología*, *18*, 47-54.

Bandrés, J., Llavona, R., & Campos, J. (1996). Luis Simarro. In M. Sáiz, & D. Sáiz (Coords.), *Personajes para una historia de la psicología en España* (pp. 185-199). Madrid: Pirámide.

Barbado, M. (1943). *Introducción a la psicología experimental*. Madrid: Consejo Superior de Investigaciones Científicas.

Bayés, R. (1972). Utilización de tórtolas en el laboratorio operante. *Revista Latinoamericana de Psicología*, *4*, 227-234.

Bayés, R. (1974). Emparejamiento (matching-to-sample) en una tórtola. *Anuario de Psicología*, *6*, 35-45.

Bayés, R. (1975). Gradiente de generalización en una tórtola. *Revista Latinoamericana de Psicología*, 7, 401-409.

30. Víctor García-Hoz Rosales, Universidad Complutense de Madrid.

Bayés, R., & Garau, A. (1982). Investigación en psicología experimental en la Universidad Autónoma de Barcelona. *Revista de Historia de la Psicología*, *3*, 73-84.

Berry, C.S. (1908). An experimental study of imitation in cats. *Journal of Comparative Neurology & Psychology*, *18*, 1-26.

Blázquez, F. (2011). A Dios por la ciencia. Teología natural en el franquismo. *Asclepio*, *LXIII*, 453-476.

Carpintero, H. (1994). *Historia de la psicología en España*. Madrid: Eudema.

Colodrón, A. (2003). Mi deuda con Pavlov. Memoria vivida en tiempo de silencio. *Revista de Historia de la Psicología*, *24*, 291-299.

Haggerty, M. E. (1909). Imitation in monkeys. *Journal of Comparative Neurology and Psychology*, *XIX*, 337-455.

Huertas, J. A., Padilla, J. M., & Montes, A. (1997). La supervivencia de la psicología en diversas instituciones madrileñas después de la guerra (1939-1953). In F. Blanco (Ed.), *Historia de la psicología española desde una perspectiva socio-institucional* (pp. 219-243). Madrid: Biblioteca Nueva.

Lafuente, E., Carpintero, H., & Ferrándiz, A. (1991). La presencia del Dr. Lafora en México. Un estudio de la psicología española en la emigración. *Revista de Historia de la Psicología*, *12*, 247-257.

de Luna, J. (1921). Notas psicobiológicas: algunas observaciones y experimentos en el ratón gris, en el albino y en el híbrido. *Archivos de Neurobiología*, *II*, 384-397.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London: Academic Press.

Mackintosh, N. J. (1983). *Conditioning and Associative Learning*. Oxford: Oxford University Press. (Spanish translation by Victoria D. Chamizo, 1988. Madrid: Editorial Alhambra.)

Marco-Igual, M. (2011). Las neurociencias y los desvaríos de la época soviética. Los médicos republicanos españoles, testigos de excepción. *Revista de Neurología*, *53*, 233-244.

Martínez, J. (2014). *El médico rojo. Vida de Juan Planelles*. Madrid: Ediciones 2010.

Martínez-Arias, A. (2009). A perspective on the development of genetics in Spain during the xx century. *The International Journal of Developmental Biology*, *53*, 1179-1191.

Mateos, A. I., & Blanco, F. (1997). La junta para ampliación de estudios e investigaciones científicas. In F. Blanco (Ed.), *Historia de la psicología española desde una perspectiva socio-Institucional* (pp. 159-200). Madrid: Biblioteca Nueva.

Morente, F. (2001). La depuración franquista del magisterio público: Un estado de la cuestión. *Hispania: Revista Española de Historia*, *208*, 661-688.

Moya, G. (1986). *Gonzalo Rodríguez Lafora: Medicina y cultura en una España en crisis*. Madrid: Publicaciones de la Universidad Autónoma de Madrid.

Muedra, V. (1948). *La perfección científica en las obras animales. Narraciones científico-recreativas (primera serie)*. Murcia: Suc. de Nogués.

Muedra, V. (1950). *Maravillas científicas en los actos animales*. Barcelona: Tip. Cat. Casals.

Muedra, V. (1955). *Ciencias naturales. Segundo curso* (4th ed.). Barcelona: Dalmau y Jover, S.A.

Otero, L. E. (Ed.) (2006). *La destrucción de la ciencia en España: Depuración universitaria en el franquismo.* Madrid: Editorial Complutense.

Pavlov, I. P. (1929). *Los reflejos condicionados: Lecciones sobre la función de los grandes hemisferios cerebrales*. Madrid: Morata. (Spanish translation of the second Russian edition of *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*.)

Planelles, J., & Luwisch, D. (1935). La acción hipoglucemiante del apetito, reflejo condicionado. *Archivos de Neurobiología*, *XVI*, 383-386.

Ramón y Cajal, S. (1921). Las sensaciones de las hormigas. *Boletín de la Real Sociedad Española de Historia Natural*, 555-572. (Special edition for the 50th anniversary of the foundation.)

Rodríguez Lafora, G., & Prados, M. (1921). Investigaciones experimentales sobre la función del cuerpo calloso. *Archivos de Neurobiología*, *II*, 1-39.

Ruiz, G., Pellón, R., & García, A. (2006). Análisis experimental de la conducta en España. *Avances en Psicología Latinoamericana*, *24*, 71-103.

Sánchez Ron, J. M. (1999). *Cincel, martillo y piedra. Historia de la ciencia en España (siglos XIX y XX)*. Madrid: Taurus.

Simón, J. (1947). *A Dios por la ciencia. Estudios científico-apologéticos* (4th ed.). Barcelona: Lumen.

Tortosa, F., Civera, C., & Esteban, C. (1998). Historia y perspectivas de la psicología en España. In F. Tortosa (Ed.), *Una historia de la psicología moderna* (pp. 531-551). Madrid: McGraw-Hill.

Tudela, P. (2010). La formación de los psicólogos en el siglo XXI. Un análisis en primera persona. *Revista de Historia de la Psicología*, *31*, 61-80.

Appendix. Publications by Professor N. J. Mackintosh in Collaboration with UB members

Diez-Chamizo, V., Sterio, D., & Mackintosh, N. J. (1985). Blocking and overshadowing between intra-maze and extra-maze cues: A test of the independence of locale and guidance learning. *Quarterly Journal of Experimental Psychology*, *37B*, 235-253.

Chamizo, V. D., & Mackintosh, N. J. (1989). Latent learning and latent inhibition in maze discriminations. *Quarterly Journal of Experimental Psychology*, *41B*, 21-31.

Trobalón, J. B., Sansa, J., Chamizo, V. D., & Mackintosh, N. J. (1991). Perceptual learning in maze discriminations. *Quarterly Journal of Experimental Psychology*, *43B*, 389-402.

Trobalón, J. B., Chamizo, V. D., & Mackintosh, N. J. (1992). Role of context in perceptual learning in maze discriminations. *Quarterly Journal of Experimental Psychology*, *44B*, 57-73.

March, J., Chamizo, V. D., & Mackintosh, N. J. (1992). Reciprocal overshadowing between intra maze and extra-maze cues. *Quarterly Journal of Experimental Psychology*, *45B*, 49-63.

Rodrigo, T., Chamizo, V. D., McLaren, I. P. L., & Mackintosh, N. J. (1994). Effects of the pre-exposure to the same or different pattern of extra-maze cues on subsequent extra-maze discrimination. *Quarterly Journal of Experimental Psychology*, *47B*, 15-26.

Sansa, J., Chamizo, V. D., & Mackintosh, N. J. (1996). Aprendizaje perceptivo en discriminaciones espaciales. *Psicológica*, *17*, 279-295.

Rodrigo, T., Chamizo, V. D., McLaren, I. P. L., & Mackintosh, N. J. (1997). Blocking in the spatial domain. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*, 110-118.

Prados, J., Chamizo, V. D., & Mackintosh, N. J. (1999). Latent inhibition and perceptual learning in a swimming pool navigation task. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*, 37-44.

Sánchez-Moreno, J., Rodrigo, T., Chamizo, V. D., & Mackintosh, N. J. (1999). Overshadowing in the spatial domain. *Animal Learning and Behavior*, *27*, 391-398.

Mackintosh, N. J., & Chamizo, V. D. (Coords.) (2002). Spatial learning and cognition (special issue). *Psicológica*, *23*, 1 (www.uv.es/psicologica).

Trobalon, J. B., Miguelez, D., McLaren, I. P. L., & Mackintosh, N. J. (2003). Intradimensional and extradimensional shifts in spatial learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *29*, 143-152.

Chamizo, V. D., Manteiga, R. D., Rodrigo, T., & Mackintosh, N. J. (2006). Competition between landmarks in spatial learning: The role of proximity to the goal. *Behavioural Processes*, *71*, 59-65.

Chamizo, V. D., Rodrigo, T., & Mackintosh, N. J. (2006). Spatial integration with rats. *Learning and Behavior*, *34*(4), 348-354.

Forcano, L., Santamaría, J., Mackintosh, N. J., & Chamizo, V. D. (2009). Single landmark learning: Sex differencies in a navigation task. *Learning and Motivation*, *40*, 46-61.

Rodríguez, C. A., Torres, A., Mackintosh, N. J., & Chamizo, V. D. (2010). Sex differences in the strategies used by rats to solve a navigation task. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*, 395-401.

Rodríguez, C. A., Chamizo, V. D., & Mackintosh, N. J. (2011). Overshadowing and blocking between landmark learning and shape learning: The importance of sex differences. *Learning and Behavior*, *39*, 324-335.

Chamizo, V. D., Rodríguez, C. A., Espinet, A., & Mackintosh, N. J. (2012). Generalization decrement and not overshadowing by associative competition among pairs of landmarks in a navigation task. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 255-265.

Rodríguez, C. A., Chamizo, V. D., & Mackintosh, N. J. (2013). Do hormonal changes that appear at the onset of puberty determine the strategies used by female rats when solving a navigation task? *Hormones and Behavior*, *64*, 122-135.

Torres, M. N., Rodríguez, C. A., Chamizo, V. D., & Mackintosh, N. J. (2014). Landmark vs. geometry learning: Explaining female rats' selective preference for a landmark. *Psicológica*, *35*, 81-100.

Chamizo, V. D., Rodríguez, C. A., Torres, I., Torres, M. N., & Mackintosh, N. J. (2014). What makes a landmark effective?: Sex differences in a navigation task. *Learning and Behavior*, *42*, 348-356.

Civile, C., Chamizo, V. D., Mackintosh, N. J., & McLaren, I. P. L. (2014). The effect of disrupting configural information on rat's performance in the Morris water maze. *Learning and Motivation*, *48*, 55-66.

COL·LECCIÓ
HOMENATGES
51

UNIVERSITAT DE
BARCELONA
Edicions