2019

Attention, Context and the Inverse Base Rate Effect

Inkster, Angus

http://hdl.handle.net/10026.1/14725

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.
This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author’s prior consent. It is licensed under the Creative Commons Attribution-Share Alike License 4.0 (Unported) (CC-BY-SA) except where stated otherwise.
Attention, Context and the Inverse Base Rate Effect

by

Angus Inkster

A thesis submitted to Plymouth University
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology
Faculty of Health and Human Sciences

July 23, 2019
Acknowledgements

First, I would primarily like to thank Andy Wills for his significant input and assistance as my supervisor. The research contained within this thesis has benefited tremendously from his expertise, and I would like to express my utmost thanks for the effort, patience and time that he has invested in me over the last three years. I have worked with Andy for the past five years and would like to think that I will continue to do so in the future.

My thanks go to Charlotte Edmunds, Chris Mitchell, Stuart Spicer, Gintaire Zaksaite and a host of others both within the School of Psychology and the Psychology PhD programme for general advice and suggestions pertaining to the content of this thesis. I am indebted to Fraser Milton for his input and assistance during the preparation and analysis of the fMRI experiment reported in Chapter 4. I would like to thank Abdelmalek Benattayallah for his technical assistance in using the fMRI machine at Exeter Medical Centre. Further thanks go to Gemma Williams, for assistance in running Experiment 1 and in the design of the cat and dog images used in this and further experiments. I would also like to thank Anna Robertson, for the creation of the cell images used in Experiment 8. My gratitude also goes to several research students who assisted in gathering experimental data.

Lastly, I would like to thank my family and friends for their support over the course of my PhD. My partner Gemma Dickinson deserves particular gratitude for her continued support and, although psychology is not her field of expertise, her advice and assistance in my efforts over the last seven years. Similar thanks go to my parents, Catherine Inkster, James Inkster and Samantha Inkster, who have continued to take interest in my work and supported me throughout my time at the University of Plymouth.
I dedicate this thesis to Gemma, and to my parents.
Author’s declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. Work submitted for this research degree at Plymouth University has not formed part of any other degree either at Plymouth University or at another establishment.

Relevant scientific conferences were regularly attended at which work was presented.

Word count for the main body of this thesis: 30,701

Signed: ________________________________

Date: _________________________________

Publications:


Presentations and conferences attended:


Abstract

The Inverse Base Rate Effect (IBRE; Medin & Edelson, 1988) is a non-rational behavioural phenomenon reported within the learning literature. A considerable amount of research has investigated the effect, as well as other effects that occur in conjunction with it. The EXIT model (Kruschke, 2001a) is widely considered to be the best formal explanation of the IBRE and is based on an error-driven attentional account. This thesis aims to further existing research, investigating unexplored avenues and providing new datasets with which to test the capabilities of the EXIT model.

Chapter 1 provides a review of the research surrounding the IBRE and explanations of the effect; identifying three potential avenues for research. First, the effect of a shift in experimental context on the IBRE and a co-occurring effect sometimes noted in IBRE procedures; the B>C effect. Second, the effect of training length on the IBRE and on responding to a novel cue. Third, neuroscientific investigation of the IBRE to identify neural activations that correlate with the effect in specific brain regions.

Chapter 2 identifies a procedure capable of producing both the IBRE and B>C effects and assesses the effect of changing the experimental context. The lack of any significant change in the magnitude of the effects renders an informal context explanation offered by Le Pelley, Mitchell, Beesley, George, and Wills (2016) inadequate, but the EXIT model successfully captures the results of these experiments. A change in responding between learned and novel contexts for a novel cue presented at test is worth noting, as this result is the opposite of previous work (Johansen, Fouquet, & Shanks, 2007; Juslin, Wennerholm, & Winman, 2001).

Chapter 3 assesses the effect of increasing training length on the IBRE and responding to a novel cue. The results of the experiments in this chapter partially replicate previous work (Johansen et al., 2007; Juslin et al., 2001), finding some evidence of an IBRE; and the pattern of responding to a novel cue at low levels of training. Of particular interest is that the response pattern for the novel cue is something the EXIT model failed to capture.

Chapter 4 reports the first functional magnetic resonance imaging study that finds a significant IBRE. Region of Interest (ROI) analysis was conducted, and informed by areas previously linked to prediction error. The ROI analysis compared the difference in activations for singular stimuli (B and C) and frequency matched controls. Activations were noted in brain areas linked to prediction error, including the caudate body, the anterior cingulate cortex and the middle frontal gyrus. Further, the analysis suggests greater activation in these areas for C compared to B, and a greater difference in activations between C and B and the frequency matched controls. This supports the error-driven learning account implemented in the EXIT model.

Chapter 5 discusses the implications of the experiments within this thesis, and how they influence further work investigating the IBRE.
Contents

Acknowledgements v
Author's declaration ix
Abstract 1

1 Introduction 11
  1.1 The Inverse Base Rate Effect 14
  1.2 Explanations of the Inverse Base Rate Effect 15
  1.3 The Le Pelley et al. (2016) Account 17
  1.4 The EXIT model 19
    1.4.1 Formal Description of EXIT 22
  1.5 Other models 27
  1.6 Structure of the Thesis 27

2 The effect of context on the IBRE and the B>C effect 29
  2.1 Introduction 29
  2.2 Experiment 1 31
    2.2.1 Method 32
    2.2.2 Results 35
    2.2.3 Modelling 37
    2.2.4 Discussion 39
  2.3 Experiment 2 41
    2.3.1 Method 41
    2.3.2 Results 43
    2.3.3 Modelling 45
    2.3.4 Discussion 46
  2.4 Experiment 3 47
    2.4.1 Method 47
    2.4.2 Results 50
    2.4.3 Discussion 52
3 The effect of training length on the IBRE and a novel cue

3.1 Introduction ................................. 61
3.2 Experiment 5 .................................. 62
   3.2.1 Method .................................. 63
   3.2.2 Results .................................. 65
   3.2.3 Modelling ................................. 68
   3.2.4 Discussion ................................. 69
3.3 Experiment 6 .................................. 70
   3.3.1 Method .................................. 70
   3.3.2 Results .................................. 71
   3.3.3 Modelling ................................. 74
   3.3.4 Discussion ................................. 75
3.4 Experiment 7 .................................. 76
   3.4.1 Method .................................. 76
   3.4.2 Results .................................. 77
   3.4.3 Modelling ................................. 79
   3.4.4 Discussion ................................. 80
3.5 Chapter Discussion .......................... 80
4 Neural Correlates of the Inverse Base Rate Effect

4.1 Introduction .................................. 83
4.2 Experiment 8 .................................. 86
   4.2.1 Method .................................. 86
   4.2.2 Results .................................. 91
   4.2.3 Discussion ................................. 96
4.3 Chapter Discussion .......................... 97
## CONTENTS

5 General Discussion

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 General Findings</td>
<td>99</td>
</tr>
<tr>
<td>5.2 The Context Explanation</td>
<td>99</td>
</tr>
<tr>
<td>5.3 Evaluation of EXIT</td>
<td>101</td>
</tr>
<tr>
<td>5.4 Neural Correlates of the IBRE</td>
<td>102</td>
</tr>
<tr>
<td>5.5 Further Directions</td>
<td>103</td>
</tr>
<tr>
<td>5.6 Final Remarks</td>
<td>105</td>
</tr>
</tbody>
</table>

References 107
List of Figures

1.1 Informal example of the context explanation ..................................... 18
1.2 The architecture of the EXIT model .................................................. 22

2.1 Novel context predictions of the context explanation ............................ 31
2.2 Context images for Experiment 1 ....................................................... 33
2.3 Trial Structure for Experiment 1 ......................................................... 34
2.4 Context image for Experiment 2 ......................................................... 42
2.5 Example trial from Experiment 3 ....................................................... 48
2.6 Trial Structure for Experiment 3 ......................................................... 49
2.7 Training accuracy in Experiment 3 .................................................... 51
2.8 Trial Structure for Experiment 4 ....................................................... 54
2.9 Training accuracy in Experiment 4 .................................................... 55

3.1 Training accuracy in Experiment 5 .................................................... 65
3.2 Proportion of rare responses to the key test stimuli of Experiment 5 ...... 67
3.3 Training accuracy in Experiment 6 .................................................... 71
3.4 Proportion of rare responses to the key test stimuli of Experiment 6 ...... 73
3.5 Training accuracy in Experiment 7 .................................................... 77
3.6 Proportion of rare responses to the key test stimuli of Experiment 7 ...... 78

4.1 An example trial presentation from the training phase of Experiment 8. . . 86
4.2 Trial Structure for Experiment 8 ......................................................... 88
4.3 Training accuracy in Experiment 8 .................................................... 91
4.4 Brain areas showing greater activation for C compared to B in Experiment 8 93
4.5 Brain areas showing greater activation for C-B compared to E-D in Experiment 8 ................................................................. 94
List of Tables

2.1 Abstract trial structure of Experiment 1 . . . . . . . . . . . . . . . . . . . . . . 32
2.2 Response proportions at test for Experiment 1 . . . . . . . . . . . . . . . . . 36
2.3 Optimal EXIT parameters for Experiment 1 . . . . . . . . . . . . . . . . . . . 39
2.4 Response proportions at test for Experiment 2 . . . . . . . . . . . . . . . . . 44
2.5 Optimal EXIT parameters for Experiment 2 . . . . . . . . . . . . . . . . . . . 45
2.6 Abstract trial structure of Experiment 3 . . . . . . . . . . . . . . . . . . . . . 48
2.7 Response proportions at test for Experiment 3 . . . . . . . . . . . . . . . . . 51
2.8 Abstract trial structure of Experiment 4 . . . . . . . . . . . . . . . . . . . . . 53
2.9 Response proportions at test for Experiment 4 . . . . . . . . . . . . . . . . . 56
2.10 Optimal EXIT parameters for Experiment 4 . . . . . . . . . . . . . . . . . 58

3.1 Abstract trial structure of Experiment 5 . . . . . . . . . . . . . . . . . . . . . 63
3.2 Response proportions at test for Experiment 5 . . . . . . . . . . . . . . . . 66
3.3 Optimal EXIT parameters for Experiment 5 . . . . . . . . . . . . . . . . . . 68
3.4 Response proportions at test for Experiment 6 . . . . . . . . . . . . . . . . . 72
3.5 Optimal EXIT parameters for Experiment 6 . . . . . . . . . . . . . . . . . . 74
3.6 Response proportions at test for Experiment 7 . . . . . . . . . . . . . . . . . 78
3.7 Optimal EXIT parameters for Experiment 7 . . . . . . . . . . . . . . . . . . 79

4.1 Abstract trial structure of Experiment 8 . . . . . . . . . . . . . . . . . . . . . 87
4.2 Response proportions at test for Experiment 8 . . . . . . . . . . . . . . . . . 92
4.3 Activation values for the whole brain analysis of C-B in Experiment 8 . . . . 94
4.4 Activation values for the whole brain analysis of (C-B)-(E-D) in Experiment 8 95
Chapter 1

Introduction

Learning is a fundamental aspect of human cognition, defined as:

“the process by which long-lasting changes occur in behavioural potential as a result of experience” (Anderson, 2000, p.4)

In an information-rich world, learning enables us to use past and present information to adapt to and overcome present and future challenges. The density of environmental information in the world on a moment-to-moment basis is excessive; processing all of it would be impractical. As such, humans have evolved to prioritise and attend selectively to important information whilst still processing other environmental information in a low-level manner. Learning contributes to this process by streamlining it; making it more efficient and reducing cognitive load. An example of this is the factory job of sexing day-old chicks. This job is considered to be difficult, given the subtle sex differences at this age. However, experts are able to perform this task with a high level of accuracy (> 98%), whilst spending less than half a second on each chick. Biederman and Shiffrar (1987) found that the accuracy of novice participants on this task increased to a level close to that of experts once they were instructed to look for a key difference that serves as an identifier of sex. This directed allocation of attention serves to highlight the particular importance of attentional learning in this task.

The above example illustrates how attending to specific information presented produces a rational judgement. However, a number of behavioural phenomena exist where humans make irrational judgements; often ignoring or not using key information to make decisions. Consider the phenomenon of base-rate neglect (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988a, 1988b; Nosofsky, Kruschke, & McKinley, 1992); of which the taxi-cab problem posed by Tversky and Kahneman (1982) is an example. In this problem,
you are told that a taxi-cab committed a hit and run accident at night. You are informed that taxi-cabs in the city are either blue or green. Two further facts are then given; that 85% of the taxi cabs are green and 15% blue, and that there was an eyewitness who, when tested with both colours, was correct 80% of the time and wrong 20% of the time. So what is the probability that the cab was blue rather than green, if the witness identified the cab as blue? Most people’s first response tends to ignore the underlying base rates of the cab colours and focuses on the fact that the witness was right 80% of the time suggesting this is the answer. However, because the base rates of the cab colours are not equal it is significantly more likely that the witness saw a green car and misidentified it as blue.

To make this a bit clearer say there are 100 cabs in the city therefore according to the facts there would be 85 green cabs and 15 blue cabs. Focusing on the 80% accuracy of the witness, these numbers are split further – 12 blue cabs are correctly identified as blue, 3 blue cabs are incorrectly identified as green, 68 green cabs are correctly identified as green and 17 green cabs are incorrectly identified as blue cabs. Taking the cases where the cab is identified as blue gives a total of 29 blue cabs, with 12 of the cabs being correctly identified. Therefore the solution to the problem is that the probability of the cab being blue based on the witness’ judgement is 12/29 or 41%; far below the 80% most people conclude.

A standard base-rate neglect procedure (Gluck & Bower, 1988b) involves participants learning to classify lists of four symptoms as one of two fictitious diseases. The occurrence of the diseases is not equal; one disease (common) occurs 75% of the time whilst the other (rare) occurs only 25% of the time. The symptoms are probabilistically related to the outcomes; for example one of the symptoms (i) has a .6 probability of occurring with the rare disease and a .2 probability of occurring with the common disease. Due to the higher occurrence of the common disease, according to Bayes’ theorem, symptom i is not predictive of either disease. Despite this, participants do not make use of the base rates of the diseases and over predict the occurrence of the rare disease based on symptom i alone. Base-rate neglect demonstrates the importance of the underlying base rates in learning and making judgements.

Another phenomenon characterised by a lack of base-rate usage is the inverse base rate effect. The Inverse Base Rate Effect (IBRE) was first reported by Medin and Edelson (1988). Previous work by Binder and Estes (1966) provided an empirical basis for the effect, but Medin and Edelson (1988) is the first published study identifying it as the IBRE.
In its canonical form participants are trained under conditions where they take the role of a medical practitioner tasked with diagnosing patients with one of two diseases; on the basis of the symptoms they present. Participants are presented with two different pairs of symptoms that can be considered abstractly using the letters AB and AC. B is perfectly predictive of the first disease and C perfectly predictive of the second disease, whilst A is uninformative. The first (common) disease occurs three times as often as the second (rare) disease. After being trained with these diseases, participants are presented with a patient who has both perfectly predictive symptoms, B and C. The rational diagnosis here would be the first, more common disease; in line with the base rates of occurrence of the two diseases. When participants are presented with a summarised form of the above information they do answer that the first disease is more likely (Johansen et al., 2007). However, when the same information is presented as a series of sequential trials the opposite occurs, with participants diagnosing patients with the second, rarer disease. This pattern of responding is called the IBRE. When presented with a patient with only the uninformative cue A, participants typically diagnose the patient with the first, common disease.

Another learning phenomenon similar to the IBRE is known as highlighting. Kruschke (2009) details highlighting in its canonical form as follows. Participants are initially trained that a compound of cues A and B leads to an outcome. As this outcome occurs early in training, it is referred to as the early outcome. Later in training participants continue to receive this compound as well as another compound; A and C. This compound leads to a different outcome; referred to as the late outcome. One important factor is that by the end of training participants have seen both outcomes equally often; and so the base rates of the outcomes are equal. Participants are then tested on cue A alone and cues B and C in compound, whilst still receiving equal AB and AC compound cues with feedback to maintain learning. The first result of interest is that participants preferentially respond with the early outcome for cue A; despite the equal base rates of the outcomes. A second result noted is that participants respond preferentially with the late outcome for the BC cue compound. These response patterns represent the highlighting phenomenon. Highlighting has been extensively investigated (Burling & Yoshida, 2017; Cunha & Laran, 2007, 2009; Kruschke, 1996a, 2009; Yoshida & Burling, 2012) and several studies have noted the similarity between highlighting and the aforementioned IBRE with Sanborn and Silva (2013) and Sewell and Lewandowsky (2012) suggesting that the phenomena are
essentially the same. Whilst they do share a high degree of similarity in terms of their response patterns, there are subtle procedural differences in how the phenomena arise; and so this thesis considers them as separate. The IBRE is the primary focus of this thesis, and is discussed in more detail in the next section.

**The Inverse Base Rate Effect**

The IBRE is a non-rational response preference occurring after learning has taken place and presents a particular challenge to purely exemplar-based accounts of learning. It can be considered as both a category learning phenomenon, where categories are constructed for further presented items to be compared to, and as a predictive learning phenomenon, due to learning to predict outcomes based on the presented stimuli. The IBRE is robust, being found in a number of studies (Bohil, Markman, & Maddox, 2005; Juslin et al., 2001; Kalish, 2001; Kruschke, 1996a; Lamberts & Kent, 2007; Medin & Bettger, 1991; Shanks, 1992; Sherman et al., 2009). Although a lot of studies make use of a medical scenario where participants diagnose patients on the basis of presented symptoms, the IBRE has also been observed in other procedural variants. Lamberts and Kent (2007) made use of pictorial stimuli; specifically pictures of objects that resembled viruses, whilst Johansen, Fouquet, and Shanks (2010) used simple shapes and lines; with both studies finding the IBRE. Kalish (2001) tasked participants with identifying an animal on the basis of a number of bars presented that they were told represented blood protein levels, again finding an IBRE. Sherman et al. (2009) found the IBRE within a procedure where participants were asked to assign fictitious people to social groups based on some presented personal traits. The ratio of disease outcome within the procedure can be manipulated and still produce the IBRE. Shanks (1992) increased the ratio of common to rare diseases to 7:1 and still found the effect. Medin and Bettger (1991) found that the IBRE could be obtained with initial 3:1 ratio training and later 2:2 ratio training, but that reversing this training method does not produce the effect.

On the subject of training, Juslin et al. (2001) followed up on a hypothesis suggested in Medin and Bettger (1991); namely that training length could reduce or remove the IBRE. Their participants completed two interleaved training and test sessions, with the results suggesting that the proportion of rare responses to the IBRE test cue reduced in the second test session relative to the first test session. A second point of interest within this study is that Juslin et al. (2001) were the first to implement a novel cue at test within an
1.2. EXPLANATIONS OF THE INVERSE BASE RATE EFFECT

IBRE procedure. The pattern of responding to the novel cue follows that of the responses to the IBRE test cue; with more rare responses to the novel cue in the first test session than to the novel cue in the second test session. However, a major issue with these results is that no statistical analysis is reported to determine if this change in responding is significant. Whilst Johansen et al. (2007) also found preferentially rare responding to a novel cue, they do not implement a training length manipulation. Don and Livesey (2017) is one of the most recent studies of the IBRE to employ a novel cue, pairing these novel cues with the shared training cue A; and interestingly found preferentially common responding. These studies represent all of the current response patterns found for novel cues within an IBRE procedure. Consequently further investigation of the novel cue would be merited; in particular the effect of variable training length.

Due to the irrationality of the effect, a considerable number of studies have attempted to model it with differing levels of success (Anderson, 1990; Gluck, 1992; Juslin et al., 2001; Kalish, 2001; Kruschke, 1996a, 2001a, 2003; Markman, 1989; Myers, Lohmeier, & Well, 1994; Shanks, 1992). Modelling the effect in isolation is simple, the problem lies in avoiding constraining a model such that it can only explain the IBRE and not other empirical data. Attempts to explain the IBRE have caused significant debate within the learning literature; and so the strengths and weaknesses of these explanations are discussed below.

Explanations of the Inverse Base Rate Effect

Historically, the first explanation of the IBRE is the relative novelty account proposed by Binder and Estes (1966). This account combines the memorable nature of novel outcomes and the idea that memorable events are predicted to be more likely to occur. Within the context of the IBRE procedure, the relative novelty account predicts that the IBRE arises as a result of cue C being more novel than cue B, due to the difference in presentation frequency during training. When presented with the BC cue at test, participants respond with the outcome associated with C; due to the novelty of the BC cue. Whilst this example attributes the novelty to the cues (symptoms), novelty can also be attributed to the outcomes (diseases), resulting in two subtly different accounts.

However, neither account is able to accommodate response patterns for other cues present in IBRE experiments. A major weakness of the relative novelty account, where novelty is attributed to cues, is that it predicts that the response pattern for the BC cue would remain the same regardless of the presence of the shared cue A at training. A number
of studies (Kruschke, 2001a; Medin & Edelson, 1988; Wills, Lavric, Hemmings, & Surrey, 2014) report that the IBRE disappears, and the response pattern reverts to following the underlying base-rate of the two diseases, if the shared symptom is not present during training. This is referred to as the shared-cue effect, and demonstrates that participants must be making use of the shared symptom, as the IBRE disappears without it. The fact that the predictions of this account are not supported by the data leaves it ineffective as an explanation of the IBRE.

The relative novelty account, where novelty is attributed to the outcomes, has a somewhat more simple weakness. When presented with the shared cue A alone, a novel occurrence, participants respond with preferentially *common* responding. If they were learning according to the relative novelty of the diseases, then participants should respond preferentially with *rare* given its less frequent occurrence. Another piece of evidence against this account is that when presented with all three abstract symptom types present in training (ABC, see Kruschke, 1996a), participants respond with preferentially *common* responding following the underlying base-rate. The predictions of this account would again suggest that participants should respond with the *rare* outcome. Regardless of what novelty is attributed to, it seems that relative novelty is inadequate as an explanation of the IBRE.

Another explanation similar to that of Binder and Estes (1966) is the eliminative inference account (Juslin et al., 2001). This account suggests that the IBRE arises as a result of what participants are able to recall about the cues and a set of rules they can apply to these recollections. Specifically, the idea is that participants remember what symptom B predicts more than what symptom C predicts because they see symptom B more often. When presented with a symptom combination of B and C, participants are more likely to remember what B predicts and forget what C predicts. However, because B and C is a novel combination of symptoms, the account suggests that they would make a novel response to a novel symptom combination. Therefore, although the disease that B predicts jumps to mind more quickly, participants respond on the basis of a rule where novelty leads to the *rare* outcome and so the response made is for the disease predicted by C. This account led to the development of the formal Eliminative Inference Model (ELMO) by Juslin et al. (2001).

ELMO is the first model to offer an explanation for novel-cue effects within the IBRE.
1.3. THE LE PELLEY ET AL. (2016) ACCOUNT

procedure. Juslin et al. (2001) found greater rare outcome responses compared to common outcome responses when participants were tested on a novel cue; a result also found in Experiment 3 of Johansen et al. (2007). ELMO has generated a significant amount of debate (Kruschke, 2001a, 2003; Winman, Wennerholm, & Juslin, 2003; Winman, Wennerholm, Juslin, & Shanks, 2005); however it suffers from the same key weakness as the relative novelty account where novelty is attributed to cues. ELMO predicts the same response pattern for the BC cue regardless of the presence of the shared cue A, something that the experimental evidence does not support (Medin & Robbins, 1971; Wills et al., 2014). Further issues for the ELMO account are found in the work of Lamberts and Kent (2007). Manipulating time pressure and cognitive load within the standard IBRE procedure showed no evidence of any effect on the magnitude of the IBRE; despite these conditions being known to adversely affect reasoning processes (De Houwer & Beckers, 2003). Finally, Kruschke (2001a) formally demonstrates that ELMO fails to capture a difference in responding to the cues A and ABC present in the behavioural data. Thus ELMO does not provide an adequate account of the IBRE.

The Le Pelley et al. (2016) Account

Le Pelley et al. (2016) offer a review of associative learning, attention and the interaction between them. Specifically they suggest that a large proportion of the data on human attentional learning can be accounted for with a simple model where attention to a stimulus is determined by its association to significant outcomes. Le Pelley et al. (2016) note that such a model would be able to capture the IBRE and the previously mentioned shared-cue effect. However, one result this model struggles with is a finding noted in some IBRE studies. In these studies, as well as testing participants on BC and A stimuli, participants are also tested on B and C individually. This, in some cases, leads to a response pattern where common disease responding to B is significantly greater than rare disease responding to C (Bohil et al., 2005; Wills et al., 2014; Winman et al., 2005). This is referred to after this point as the B>C effect. The co-occurrence of the IBRE and B>C effect is a major problem for the model proposed by Le Pelley et al. (2016). In short, the paradoxical occurrence of a situation where C attracts more attention (IBRE) but has lower associative strength than B (B>C effect) is something the model cannot reconcile due to its equation of attention and associative strength.

Le Pelley et al. (2016) do offer a potential reconciliation for this failing, appealing to the role
1.3. THE LE PELLEY ET AL. (2016) ACCOUNT

of experimental context. This informal context explanation suggests that the procedural context of the experiment itself can be considered as a cue. Considering the classic IBRE medical diagnosis paradigm, the experimental context could be the role the participant is asked to take on, the location where they are making diagnoses or the patients that they are diagnosing. Experimental context is present in all behavioural experiments in some way, but most experiments equate outcome frequency. In a procedure such as the IBRE however, the outcomes are not equally frequent. If the context is then treated as a cue then it would become associated with the outcomes and given the more frequent common disease outcome in the IBRE procedure, it would be more associated with the common outcome than the rare one. If this is assumed, then a situation where both the IBRE and B>C effect occur could be explained.

At this point, it is worth providing a more detailed example of the informal context explanation, and how specifically the experimental context might produce the associative strength advantage for B over C. The orange bars in Figure 1.1 show some assumed associative strengths for the training cues at the end of training in the IBRE procedure. These associative strengths are presented on a scale running from 5 to -5, where 5 represents perfect rare outcome responding and -5 indicates perfect common outcome responding. X represents the context cue and its ability to gain some associative strength.

![Figure 1.1: Informal example of the context explanation of both the IBRE and B > C effect. Orange bars represent assumed learned associative strengths of individual cues at the end of training. Blue bars represent predicted associative strengths for test stimuli.](image_url)
C is the only perfect predictor of the *rare* outcome, and so it gains a lot of associative strength for the *rare* outcome. B is the only perfect predictor of the *common* outcome and so it would gain associative strength for the *common* outcome. Both B and C occur alongside the A and X cues. However, as the frequency of the *common* outcome is greater than that of the *rare* outcome, cues X and A gain more associative strength to the *common* outcome. As a result of this, this explanation assumes that B competes with both A and X for attention on *common* outcome trials, and so its associative strength to the *common* outcome is hampered.

Looking at the cues presented at test, if we assume that the context remains the same and simply sum associative strengths for compound cues and include the influence of the context itself, then both the IBRE and B>C effect emerge. The blue bars in Figure 1.1 demonstrate this; e.g. $XBC = -2(X) - 2(B) + 5(C) = 1$. This suggests that the BC cue would lead to the *rare* outcome, an IBRE, and that the *common* outcome for B would be greater than the *rare* outcome for C. A is assumed to be equivalent to B at test (i.e $XA = XB$) and so would lead to the *common* outcome. A novel cue, N, is also included in this explanation as a test for the associative strength of the context; given that participants should have no associative strength for N alone. According to the informal context explanation, N should lead to the *common* outcome. Previous data on responding to a novel cue in an IBRE procedure (Johansen et al., 2007; Juslin et al., 2001) suggests the opposite, with N leading to the *rare* outcome. However, given the lack of statistical analysis in these studies it is hard to judge the validity of this result, and so the possibility remains that the predictions made by the context explanation are correct. Whilst the context explanation has some logical merit, an experimental investigation of the explanation is yet to be conducted.

**The EXIT model**

The EXIT model is the final product of a series of model developments originating in the work of Kruschke (1992). Kruschke presented the Attention Learning COVEring map (ALCOVE); a hybrid model combining the exemplar modelling of Nosofsky (1986) with the connectionist networking of Gluck and Bower (1988a, 1988b), augmented such that dimensional attentional strengths are learned gradually through gradient descent upon error. ALCOVE’s key strength was its ability to capture a range of phenomena; particularly studies of base-rate neglect (Kruschke, 1992; Nosofsky et al., 1992). However,
Lewandowsky (1995) suggested that the model’s ability to explain base-rate neglect was down to a methodological flaw in the original study (Estes et al., 1989). Specifically, in the Estes et al. (1989) study, and in the subsequent simulations of ALCOVE, a particular set sequence of 240 training trials were used. Lewandowsky (1995) demonstrated that the way in which ALCOVE accounted for base-rate neglect was directly tied to this sequence, and when this sequence was altered ALCOVE was unable to capture base-rate neglect. A further issue with ALCOVE is the use of a single set of attention strengths for all of the exemplars represented within the model; leaving it unable to account for the IBRE (Kruschke, 1992; Nosofsky et al., 1992).

The failures of ALCOVE served as a starting point for the development of a model specifically designed to capture the IBRE, the Attention to Distinctive Input (ADIT) model (Kruschke, 1996a). ADIT is similar to ALCOVE in that both models are error-driven network models, however differs in that there is no exemplar representation. Instead, ADIT implemented individual attentional strengths for the cues represented within the model, allowed the influence of base-rate knowledge and contained a mechanism for attentional shift. Such a mechanism has some historical precedent in the work of Mackintosh (1975), but Kruschke (1996a) details the first application of this idea to the IBRE.

Attentional shift is a phenomenon linked to error-driven attentional theories of learning which offer their own account of the IBRE. Due to the more frequent occurrence of common disease within the IBRE procedure, a logical assumption is that the outcome associated with the common disease is learnt before the outcome associated with the rare disease. Error-driven attention suggests that during learning participants endeavour to reduce the number of errors they make as quickly as possible. Participants learn about both A and B in terms of them leading to the common disease. This then results in an attentional shift when presented with A and C together, with attention being directed away from A and towards C to promote new learning and avoid error; in this case responding with the common disease on the basis of A. A and B are in competition in terms of learning the association to the common diseases, and so whilst both gain associative strength with the common disease they do not reach the same level of associative strength as the association of C to the rare disease. This then drives the IBRE, as when presented with B and C in combination the attention to C is greater than the attention to B. A further advantage of this account is that it accommodates the noted base-rate response patterns for A presented alone and the shared-cue effect. ADIT captures the response pattern for
A presented alone by learning that the A cue is associated with the common outcome. For the shared-cue effect consider the case where the model is given two compounds that do not share a cue i.e. DF leading to the common outcome and EG leading to the rare outcome. When tested on D and E in compound, as attention does not shift away from a shared cue in training, the E cue does not gain an advantage in terms of associative strength. As a result of this, the responding reverts back to representing the underlying base rates for the outcomes; the shared-cue effect.

ADIT’s attention shift mechanism promotes the learning of new associations whilst maintaining those associations already learned. The mechanism shifts attention away from cues in response to error, before associative weights are adjusted. ADIT has been fit to several datasets successfully (Fagot, Kruschke, Depy, & Vauclair, 1998; Kruschke, 1996a). However, one issue with the attentional shift mechanism is that attention shift is reset at the start of every trial supplied to the ADIT model. This lack of learning about attention itself led to the development of the Extended ADIT (EXIT) model (Kruschke, 2001a) to address this issue. EXIT differs from ADIT in its ability to learn about attentional reallocation, but retains the ability to capture the results occurring in conjunction with the IBRE. It further allows the representation of exemplars, and implements exemplar-based attention. EXIT has met with some criticism (Winman et al., 2003) suggesting that the model and the error-driven attentional account it represents are unable to accommodate the previously mentioned B>C effect (Bohil et al., 2005; Wills et al., 2014; Winman et al., 2005). Kruschke (2003) acknowledged this criticism, but gave a method for capturing the effect. This required the implementation of a bias cue representing the underlying base-rate of outcomes. The model achieves this representation by assuming the cue is present on every trial, and so encodes all of the outcomes that occur within an experiment. If one outcome occurs more frequently, then the bias cue becomes more associated with that outcome. As well as this, Kruschke (2003) weighted the response pattern associated with the B>C effect more heavily during model fitting, which allowed it to be captured by EXIT.

Further studies support the error-driven attention account implemented within EXIT, including eye-tracking (Kruschke, Kappenman, & Hetrick, 2005a) and Electroencephalography (EEG) (Wills et al., 2014) data. The work of Wills et al. (2014) as well as other neuroscience methodologies investigating the IBRE (O’Bryan, Worthy, Livesey, & Davis, 2018) will be discussed in more detail in Chapter 4 of this thesis. EXIT has been applied to other
1.4. THE EXIT MODEL

learning phenomena analogous to the IBRE including highlighting (Kruschke, 1996a, 2009) and illusory correlation (Sherman et al., 2009); although its account of illusory correlation has met criticism (Kutzner & Fiedler, 2015). EXIT continues to influence further research, including how the model handles missing stimulus information (Wood & Blair, 2011) and how it might apply to the formation of impressions about groups (Hegarty & Bruckmüller, 2013; Huang, Sacchi, & Sherman, 2017). Overall, the EXIT model represents the best formal explanation of the IBRE within the current literature, with recent work supporting this position (Don, 2018). Due to this, the behavioural results of the experiments within this thesis are simulated using the EXIT model; and so what follows is a full description of the model’s formal implementation.

Formal Description of EXIT

Figure 1.2 shows the architecture of the EXIT model (Kruschke, 2001b, p. 1389). Part (A) shows that for each input cue there is a corresponding input node, and for each outcome a corresponding output node. If cue $i$ is present on a trial then the activation of input node $a_{in}^i$, is set to 1. If cue $i$ is missing, then the activation is set to 0. Similarly, if outcome

![Figure 1.2: The architecture of the EXIT model as seen in Kruschke (2001b). This figure demonstrates the model in the case of two cues, one exemplar and two outcomes. (A) represents the links from cues to outcomes, with the thicker lines representing learnable associative weights. (B) represents the mechanism of allocating attention to cues. The criss-crossing lines represent the normalisation of attention. The X’s in boxes represent the multiplicative nature of attention on the cues. (C) represents the mechanism for learning attentional distributions, with thicker arrows again representing learned associative weights. Finally, (D) represents the full model, comprising of (A), (B) and (C). From "Toward a Unified Model of Attention in Associative Learning" by J. K. Kruschke, 2001, Journal of Mathematical Psychology, 45, p. 816. Copyright 2001 by Elsevier Science.](image-url)
1.4. THE EXIT MODEL

If an outcome occurs, then output node $k$ receives a teaching signal, $t_k = 1$, indicating that the node should be activated. If the outcome is absent, then $t_k = 0$. The links between the input nodes and output nodes denote associative weights; for the above nodes $i$ and $k$ the associative weight between them is represented as $w_{ki}$.

Upon stimulus presentation, the corresponding input nodes activate and activation spreads to the output nodes via the weighted connections. At this point, input activation is modulated by the attentional weights as can be seen in part (B) of Figure 1.2. Activation of output node $k$ is determined by a weighted sum across the attentionally modulated input nodes such that:

$$ d_{k}^{\text{out}} = \sum_i w_{ki} \alpha_i a_i^{\text{in}}, $$

(1.1)

$\alpha_i$ represents the attention strength on input node $i$. How these attentional strengths are determined will be discussed later. The values for $w_{ki}$ are initially 0, and change as a result of learning.

Category node activations are mapped to response probabilities such that the probability of choosing a category, $c$:

$$ p(c) = \frac{\exp(\phi a_{c}^{\text{out}})}{\sum_k \exp(\phi a_{k}^{\text{out}})}, $$

(1.2)

where $\phi$ represents a scaling constant. The probability of classifying a stimulus into category $c$ is determined by the magnitude of activation of category $c$ compared to the sum of all category activations. $\phi$ is representative of how decisive the model is, with higher values representing a higher degree of decisiveness.

Parts (B) and (C) of Figure 1.2 show the attentional sub-system of the model. The attentional sub-system maps input activations to attentional strengths. An important assumption the model makes for the attentional sub-system is that there is a finite limit to its capacity. Practically this means that increased attention to one cue reduces attention to the other cues. This is formally represented within the model by assuming that each cue has an inherent attentional gain value that is normalised to generate attentional strengths. Part (B) shows this through the criss-crossing connections between the gain and attention nodes.
The role of the attentional sub-system is twofold; to implement the assumption that any presented cue gains some attention and to learn how to distribute attention to cues presented in combination. The first goal is accomplished through each cue having a connection to a corresponding gain node, as can be seen in part (B). The second goal is accomplished by implementing exemplar mediation of the mapping from input to output. Practically, this is achieved by the model recruiting exemplar nodes when encountering a novel combination of cues. An example of this can be seen in part (C). The connection weights from exemplar nodes to gain nodes adjust such that the model can learn to predict the shift in attentional gain.

The activation of an exemplar node is determined through comparison of its similarity to a presented stimulus and the exemplar represented by the node. Similarity drops off exponentially as a result of increased distance in psychological space, which is computed using a city-block metric for psychologically separable stimulus dimensions. The formal definition for the activation of exemplar node $x$ is:

$$a_{ex} = \exp \left( -c \sum \sigma_i |\psi_{xi} - a_{in}^i| \right), \quad (1.3)$$

The constant $c$ represents the specificity; i.e. how close in similarity a stimulus has to be to the exemplar represented by a specific node in order for that node to activate. $\psi_{xi}$ is indicative of the presence or absence of cue $i$ in exemplar $x$. If the cue is present then $\psi_{xi} = 1$ and if absent then $\psi_{xi} = 0$.

Attention propagates to the gain nodes from the input nodes through two paths in the attention sub-system; that can be seen in parts (B) and (C). The activation of the gain node $i$ is determined by:

$$g_i = a_{in}^i \sigma_i \exp \left( \sum_x w_{ix} a_{ex}^x \right), \quad (1.4)$$

where $w_{ix}$ represents the associative weight between exemplar node $x$ and gain node $i$. These weights are again initialised at 0, and change as a result of learning. By design, zero gain is given to input cues with 0 activation. If a cue the model has not seen before is presented, then the activation of the gain node is set to 1. This is achieved through the exponential within Eq. 1.4, as the weight between the gain node for that cue and the exemplars would be 0, and so the exponential would evaluate to 1. One final point to note
1.4. THE EXIT MODEL

is that all gains on cues are non-negative due to the use of an exponential.

ADIT originally used a separate formula for mixing the base rates of categories with the choice probabilities generated. Instead, EXIT uses the previously mentioned bias cue (Kruschke, 2003) to represent this. The salience of the bias cue differs in respect to the other cues. This is implemented within EXIT through use of a parameter $\sigma$, present in Eqs. 1.3 and 1.4. Whilst all cues have a $\sigma$ value in these equations, these values for the presented cues are all set to 1. The value of the parameter for the bias cue can be set such that $0 < \sigma < 1$. The closer $\sigma$ gets to 1 the more influence the bias cue has on responding, and conversely the closer to $\sigma$ gets 0 the less influence it has.

Activation then propagates from the gain nodes to the attention nodes. The limit to the capacity of the attention sub-system is implemented through a Minkowski power metric. This limit is denoted by $\sum \alpha_i^p = 1$, with $P$ representing the power within the Minkowski metric. Attention to cue $i$ is calculated through normalising the gain of cue $i$ as follows:

$$\alpha_i = \frac{g_i}{\left( \sum_j g_j^p \right)^{1/P}}, \quad (1.5)$$

The denominator will always be greater than 0, as the gains calculated will always be non-negative and at least one will be $> 0$. Increased attentional capacity is reflected in greater values of $P$. If $P = 1$ then the attention to a single cue is just the relative proportion of its gain against the sum of gains to all cues. Increasing attention to one cue causes an equal decrease in attention to all the other cues. As $P$ gets significantly larger, the cue with maximal gain gets an attentional strength of close to 1, with other cues getting attentional strengths proportional to the maximal gain. In this case, several cues can get attentional strengths of close to 1, reflecting a high degree of attention paid to the cues presented. If $0 < P < 1$, then an increase in attention paid to one cue results in a decrease of attention to all other cues. In this case, attention is very limited with attention only being paid to a cue if it is the only one to not have an attention strength of close to 0.

After the model determines the category node activations and the probabilities of making each categorisation response, it receives corrective feedback. The model prioritises the reduction of error through rapid attention shifting. This attention shift is achieved through the implementation of gradient descent on error and calculated according to the following
1.4. THE EXIT MODEL

\[
\Delta g_I = \lambda_g \sum_k (t_k - a_k^{\text{out}})(w_k a_I^{\text{in}} - a_k^{P-1} a_k^{\text{out}}) / \left( \sum_j g_j^P \right)^{1/P},
\]

(1.6)

\(\lambda_g\) represents a positive constant defined as the shift rate of attention. As attention is hypothesised to shift a large amount on a single trial, it follows that this needs to occur formally within the model. However, achieving this in a single step is not possible, due to the non-linear nature of gain. In simpler terms, as attention changes, the gradient for gain changes as well. To solve this the change generated by the gain change equation is iterated 10 times, to approximate the non-linearity. After each iteration, activation is repropagated to generate a new error. If at any point the gain value is driven negative, then the value is reset to 0 before the next iteration.

After attention shifts the association weights are then adjusted, again by gradient descent on error as follows:

\[
\Delta w_{KI} = \lambda_w (t_K - a_K^{\text{out}}) a_I^{\text{in}},
\]

(1.7)

where \(\lambda_w\) is a constant representing the learning rate for output weights.

Finally the associative weights for the gain nodes are also adjusted by gradient descent on error. Error here is defined as the sum of squared differences between the shifted value and the pre-shift value. This is represented formally as:

\[
\Delta w_{gI} = \lambda_x (s_I^{\text{shift}} - s_I^{\text{init}}) a_I^{\text{ex}},
\]

(1.8)

where \(\lambda_x\) is the constant for learning rate for the associative weights between the exemplar and gain nodes.

The full formal implementation of EXIT therefore contains seven free parameters:

- the response probability scaling constant, \(\phi\), in Eq. 1.2
- the specificity, \(c\), of the exemplar nodes in Eq. 1.3
- the attentional normalisation power, \(P\), in Eq. 1.5

*For a full description of the derivation of this equation, see Kruschke (2001b).
• the attention shift rate, \( \lambda_g \), in Eq. 1.6

• the associative weight learning rate, \( \lambda_w \), in Eq. 1.7

• the learning rate, \( \lambda_x \), for the associative weights from exemplar to gain nodes in Eq. 1.8

• the salience of the bias cue, \( \sigma \), that determines the influence of the bias cue on responding, in Eq. 1.3 and Eq. 1.4.

Other models

Further explanations of the IBRE are also based on the work of Gluck and Bower (1988a); Gluck (1992)’s adaptation and Shanks (1992)’s Attentional Connectionist Model (ACM). Whilst these models offer an account of the IBRE this thesis focuses exclusively on the EXIT model. This is due to its influential status within the literature; specifically the considerable amount of work citing the model.

Structure of the Thesis

The main purpose of this thesis is to provide further insight into the IBRE and the effects that occur in conjunction with it. Chapter 2 aims to provide the first experimental examination of the informal context explanation offered by Le Pelley et al. (2016). Over the course of four experiments, a procedure that can produce the IBRE and B>C effect is identified. The experimental context is then manipulated to assess the effect of changing context on these effects. The EXIT model is fit to each of the experiments containing a context manipulation within this chapter; to evaluate its ability to capture the behavioural results.

Chapter 3 focuses on the effect of manipulating training length, as reported in Juslin et al. (2001). In Juslin et al. (2001) the response patterns for the BC cue and a novel cue were preferentially rare after a single training session, and the magnitude of this preference reduced for the BC cue and the novel cue after a second training session. No statistical analysis is reported for these results, and so over three experiments the response patterns for the BC cue and the novel cue are examined at low and high levels of training under within-subject and between-subject conditions to assess the veracity of the results reported in Juslin et al. (2001). EXIT is again fit to the behavioural results of these experiments to further assess its ability to capture behavioural data.

Chapter 4 reports the first known instance of an experiment that finds a significant IBRE
under an fMRI methodology. Whilst O’Bryan et al. (2018) report an IBRE under an fMRI methodology, an inherent flaw within their analysis casts doubt on the veracity of their findings. Using the procedure identified in Chapter 2 that produces an IBRE, and making use of an analysis method used successfully in Wills et al. (2014), neural correlates underpinning the effect are identified through Region of Interest (ROI) analysis.

Finally, Chapter 5 serves to draw the findings of these chapters together, discussing the implications of this work in terms of the explanations it supports, and the future directions that could be taken to further investigate the IBRE.
Chapter 2

The effect of context on the IBRE and the B>C effect

Introduction

A result sometimes noted in the IBRE literature is the apparent paradox of both an IBRE and B>C effect occurring in conjunction with each other (Bohil et al., 2005; Medin & Edelson, 1988; Wills et al., 2014; Winman et al., 2005). This result is unpredictable, occurring in studies with different procedures, but troubling given that it suggests both an attentional advantage for C over B (IBRE) and a greater associative strength for B over C (B>C). Le Pelley et al. (2016) offer a potential explanation of this. Specifically they suggest that the experimental context exists as a cue present on every trial. The context cue becomes associated with the outcomes of the experiment, much like the other cues. Whilst this is irrelevant for studies where the outcome frequency is matched, for the IBRE procedure the context would be more strongly associated with the common outcome. This explanation lacks a formal investigation; but Figure 1.1 shows an informal example of how the context explanation might produce both the IBRE and B>C effect in conjunction through some assumed learned associative strengths. The cue representing context, gains associative strength for the common outcome. The associative strengths for the cue compounds presented at test are summed to determine the proportion of predicted common or rare outcome responses made. As a result the XBC cue compound leads to slight rare responding, as well as more common responding to the XB cue compound compared to rare responding to the XC cue compound. As well as these response patterns, the context explanation suggests that the XN cue compound, where N is a novel cue, would produce common responding.

This is particularly striking as both Juslin et al. (2001) and Johansen et al. (2007), previous studies that employed a novel cue at test, show preferentially rare responding to a novel
2.1. INTRODUCTION

cue. The response pattern in these studies is purely ordinal; no statistical analysis is reported. The singular case of preferentially common responding to a novel cue is reported in Don and Livesey (2017). However, this result should be interpreted with caution; due to the fact that their novel cue was presented in compound with the shared cue, A, present in training. A has been found to produce preferentially common responding repeatedly (Kruschke, 1996a, 2001a; Wills et al., 2014), so it is likely that responding according to the presence of A drives the pattern of responding reported in Don and Livesey (2017). Given the low number of studies employing a novel cue, the possibility remains that the predicted novel cue response pattern of the context explanation is correct.

If we assume that the response patterns predicted by the context explanation are correct, then we can predict what the change in these response patterns would be if the experimental context were altered. The veracity of these predictions could then be tested experimentally. Figure 2.1 shows how these response patterns would change if test occurred under a novel context. The novel context Y has no associative strength to either the common or rare outcomes. As a result of this, the cue compound YBC would produce a larger proportion of rare responding than the XBC cue compound, indicating a larger IBRE. The B>C effect is predicted to reverse, with more rare responding to the YC cue compound than common responding to the YB cue compound. As A and B are assumed to be equivalent, common responding to A should decrease. Finally, the YN cue compound should produce random responding.

Chapter 1 concluded that the formal model currently most able to explain results around the IBRE is the EXIT model (Kruschke, 2001a). Interestingly, the model contains a mechanism that operates in a similar way to the context cue suggested by the context explanation. Specifically, EXIT implements a bias cue, which is present on every trial in the experiment. This bias cue in an IBRE experiment would become associated with the common outcome due to the difference in outcome frequency, in much the same way as the context cue. EXIT has previously been shown to capture both the IBRE and B>C effect (Kruschke, 2001a, 2003), however no simulation published has tasked it with capturing novel cue response patterns.

The aims of this chapter are threefold: one; to identify a procedure capable of producing both the IBRE and B>C effect in conjunction, two; to implement a context manipulation within an IBRE procedure to evaluate the predictions made for a novel test context by the
2.2. EXPERIMENT 1

Figure 2.1: A demonstration of how the IBRE, B>C effect and response to a novel cue would change under a novel context.

Context explanation and three; to evaluate the ability of the EXIT model to capture the behavioural data of these experiments; in particular the response patterns to the novel cue.

Experiment 1

Experiment 1 aims to implement a procedure that both produces the IBRE and B>C effect and contains a context shift to a novel context at test. Given that the standard IBRE procedure is a medical diagnosis task, this was used as the basis for the procedure of Experiment 1. However, in order to implement a context shift, a key aspect of the task was changed. Instead of diagnosing human patients according to symptoms, the task instead involved diagnosing cats or dogs. This change was primarily to facilitate the change in experimental context, which was represented by the specific animal being diagnosed. For example, if someone were to be trained to diagnose cats, they would then be asked to diagnose cats (same-context) and dogs (different-context) at test. Criterion-based training was used to ensure an appropriate level of learning before participants were tested. The key outcomes for this experiment are to produce the IBRE and B>C effect in conjunction, to examine the extent to which the behavioural response patterns support the predictions of the context explanation and to assess the ability of the EXIT model to capture the behavioural data patterns.
2.2. EXPERIMENT 1

Method

Participants

48 people participated in this experiment. In all the experiments of this thesis, participants were either undergraduate students at the University of Plymouth or University of Exeter who completed the experiments for partial fulfilment of course credit/monetary recompense, or members of the general public who completed the experiments for monetary recompense. Participants did not complete more than one of the experiments and were a range of ages, genders and ethnicities. Specific demographic information was not recorded.

Apparatus

For all experiments in this chapter, visual stimuli were displayed on 22 inch, 60 Hz Philips Monitors using standard PCs and keyboards. The distance between the participant and the screen in these experiments was approximately 50cm. Behringer HPM1000 headphones were used for the presentation of auditory information. The experiment was constructed using the PsychoPy experimental software (Peirce, 2007), version 1.83.04.

Stimuli

Symptoms and diseases were presented in black text against a white background. Symptoms applicable to either a cat or dog were used: chesty cough, fur loss, weepy eyes, vomiting, fatigue, diarrhoea, bad breath or skin ulcers. Four fictional disease names were also used: Fips Syndrome, Gertz Syndrome, Haust Syndrome and Jominy Syndrome. Table 2.1 shows the combinations of abstract cues and diseases presented in the training and test phases of the experiment. The cue N is a novel cue presented only at test: N₁ appears in the first test phase, and N₂ in the second. The symptom paired with each abstract cue was randomised for each participant. Diseases were mapped to the abstract

<table>
<thead>
<tr>
<th>Training trials (relative frequency)</th>
<th>Test trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1 B_1 \rightarrow \text{common}_1$ (x 3)</td>
<td>$A_1, A_2, B_1, B_2, C_1, C_2$</td>
</tr>
<tr>
<td>$A_1 C_1 \rightarrow \text{rare}_1$ (x 1)</td>
<td>$AB_1, AB_2, AC_1, AC_2$, $BC_1, BC_2, ABC_1, ABC_2$, $N_{1/2}$</td>
</tr>
<tr>
<td>$A_2 B_2 \rightarrow \text{common}_2$ (x 3)</td>
<td>x 6</td>
</tr>
<tr>
<td>$A_2 C_2 \rightarrow \text{rare}_2$ (x 1)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Abstract trial types for the training and test phases of Experiment 1. Bold type highlights the test stimuli of particular theoretical interest.
2.2. EXPERIMENT 1

Figure 2.2: The two context images that stimuli cues were presented on for Experiment 1.
Image author: Gemma Williams, CC-BY-SA 4.0

disease types in two different ways across participants.*

Symptoms and diseases were presented in the context of either diagnosing a cat or a
dog. To emphasise the context, the symptoms were presented on the flank of the relevant
animal. Figure 2.2 shows the images used and how the symptoms were presented. The
images were 980 x 600 pixels in size. Animal-relevant noises were also used to further
emphasise the context (six for each species, randomly selected on each trial, each sound
1000–2000 ms in duration).

Procedure

Participants received a set of initial instructions, asking them to take on the role of a
fictitious veterinarian, diagnosing either cats or dogs according to their physical symptoms.
Once they had read the instructions, they proceeded to the training phase. Figure 2.3
shows the structure of a trial in Experiment 1. Each trial in the training phase started
with presentation of the cat or dog image and sound. After 400 milliseconds (ms), the
symptoms appeared as a vertical list on the flank of the animal. The order in which
symptoms appeared on the vertical list was counterbalanced across trials; and trial order
was also randomised.

Participants diagnosed the animal by pressing one of four keys: “f” for Fips Syndrome,
“g” for Gertz Syndrome, “h” for Haust Syndrome and “j” for Jominy Syndrome. After
making a response, participants were given corrective feedback for 2000 ms, telling them
whether they were correct or incorrect and also what the correct response key and disease
were. This feedback was presented on the animal's flank, replacing the symptoms. If

*Either: (1) Fips → common₁, Gertz → rare₁, Haust → common₂, Jominy → rare₂, or (2) Fips → rare₁, Gertz →
common₁, Haust → rare₂, Jominy → common₂
2.2. EXPERIMENT 1

Figure 2.3: The structure of a trial in Experiment 1. The context is displayed from 0 - 30400ms. The symptom(s) is/are displayed from 400 - 30400ms. The feedback is displayed from 30400 - 32400ms. If a response was made while the context and symptom was on screen, then the context and symptom disappeared, and feedback was displayed for 2000ms from that point.

the participant did not respond within 30 s, a time-out message appeared for 2000 ms, informing them they were out of time and should speed up.

After each training block of 32 trials, participants were informed of the percentage of trials they had diagnosed correctly on that block, and told that they should aim to achieve 100% correct. Training was to criterion, with participants’ accuracy assessed under a separate criterion for common stimuli (AB) and rare stimuli (AC). If participants did not score at least 21/24 on the common stimuli and at least 7/8 on the rare stimuli, they received an additional block of training. Once a participant either met this criterion or received a total of 15 training blocks, they were moved onto the first test phase.

Before the first test phase began, participants received further instructions. They were informed that they would now be presented with combinations of symptoms that they had not seen before, without corrective feedback. They were also told which species they would be diagnosing. The trial structure of the test phase was identical to the training phase except that, after making a response, instead of receiving corrective feedback, participants now received the message “Response recorded” for 2000 ms. Participants received 90 test trials, and then proceeded to the second test phase.

Before the second test phase began, participants received further instructions, informing them that they would still not receive feedback. They were also told which species they would be diagnosing. The second test phase then proceeded in the same manner as the first test phase. After participants completed 90 trials in this phase, the experiment finished.
Results

Statistical analysis for all of the experiments within this thesis was conducted within the R environment (R Core Team, 2018), with additional packages BayesFactor (Morey & Rouder, 2018), ez (Lawrence, 2016), and psych (Revelle, 2018). All NHST tests were conducted at an alpha level of .05. Bayes Factors less than one third were interpreted as substantial evidence for the null, whilst those greater than 3 were interpreted as substantial evidence for the alternative. Parametric tests were conducted throughout this thesis. The experimental data was checked for normality, and where the distribution was non-normal, non-parametric tests were conducted. These tests were not substantively different from the parametric tests and so for consistency only the parametric tests are reported.

The trial-level raw data, analysis scripts and modelling scripts for this experiment will be available at https://osf.io/7eqgw/ once this experiment is published. Due to the experiment having two contexts (cats or dogs), and two test phases, there were four counterbalance conditions: CCD (training: cats, test 1: cats, test 2: dogs), CDC, DDC and DCD. The conditions that are abstractly the same (CCD/DDC and CDC/DCD) were combined resulting in two conditions with 26 and 22 participants in each condition respectively.

All participants reached criterion, taking on average 2.77 (SD = 1.40) blocks to do so. A three-way ANOVA was conducted on the training phase data, looking at the effects of training block (first/last), stimulus frequency (common/rare), and training context (cat/dog) on accuracy. Accuracy in the final block was significantly higher than accuracy in the first block, $F(1,46) = 129.11, p < .001$. Participants were significantly more accurate with the common stimuli than the rare stimuli, $F(1,46) = 10.19, p = .003$. Participants who were trained under the cat context were more accurate than those trained under the dog context, $F(1,46) = 6.47, p = .01$. A significant interaction was found between stimulus frequency and block, $F(1,46) = 4.68, p = .04$, and training context and block, $F(1,46) = 6.04, p = .02$, reflecting the fact that the effects of stimulus frequency and context diminished as training proceeded. No other main effects or interactions were significant. A two-way ANOVA on the final block of training shows that participants were significantly more accurate on the common stimuli than the rare stimuli, $F(1,46) = 4.30, p = .04$. No other significant main effects or interactions were found.

In the test phase, all participants received both a same-context test phase (where the
species being diagnosed was the same as in training), and a different-context test phase (where the species was different). The order of these test phases was counterbalanced across participants. In the following analyses, as well as the analyses for the other experiments employing context shift in this chapter, context (same vs. different) is treated as a within-subjects factor. Table 2.2 shows the response proportions for each stimulus under both different-context and same-context conditions. Cues that are abstractly identical have been combined in this Table. For example, “A” represents responses to both $A_1$ and $A_2$. Responses in this Table are classified as common, rare, common-other and rare-other. Common and rare denote the common and rare diseases with which the cue was paired during training; common-other and rare-other denote the common and rare diseases with which the cue was never paired. Due to the low probability of common-other and rare-other responses, these response types were excluded from further analysis. Specifically, for the following analyses the response proportions were rescaled such that the common and rare response proportions within each condition summed to 1.

In the following analyses, data is first analysed within the same and different conditions, and then between the two conditions. All of the within-subject t-test analyses are compared against a level of 0.5 representing random responding.  

Looking first at the same-context condition, rare responses to the BC stimulus were found to be significantly higher than .5, $BF = 34.27, t(47) = 3.57, p = .001$, indicating the presence of an IBRE. A t-test between the common response proportion to the B stimulus and

† It is worth noting that in general, random responding when there are four response options is set to the 0.25 level. However, in the case of this experiment, as the common-other and rare-other response options are removed, the level is set to 0.5.
the rare response proportion to the C stimulus suggests that rare responses to C were significantly higher, $BF = 1.61, t(47) = 2.27, p = .028$. This is the opposite of what was predicted; although the Bayes is inconclusive. The proportion of common responses to the A stimulus was significantly greater than .5, $t(47) = 4.76, p < .001$. Common responses to the novel N stimulus were significantly greater than .5 $BF = 1.85, t(47) = 2.34, p = .023$; although again the Bayes is inconclusive.

Looking second at the different-context condition, rare responses to the BC stimulus were also found to be significantly higher than .5, $BF = 13.77, t(47) = 3.23, p = .002$, again indicating the presence of an IBRE. Rare responses to the C stimulus were significantly higher than common responses to the B stimulus, $BF = 2.71, t(47) = 2.53, p = .015$, although again the Bayes is inconclusive. The proportion of common responses to the A stimulus was significantly greater than .5, $t(47) = 6.82, p < .001$. Common responses to the novel N stimulus did not differ from .5, $BF = 0.28, t(47) = 1.10, p = .28$.

Looking third at the difference in response proportions between conditions, there was no difference in the proportion of rare responses between conditions for BC stimuli, $BF = .21, t(47) = .80, p = .43$. Similarly there was no difference between the conditions for common responses to the B stimulus, $BF = .17, t(47) = .42, p = .68$, or rare responses to the C stimulus, $BF = .18, t(47) = .52, p = .61$. There was significantly greater common responding in the different-context compared to the same-context condition for the A stimulus, $t(47) = 2.22, p = .032$. There was evidence of an effect of context shift on the novel N stimulus, with greater common responding in the same-context condition than the different context condition, $BF = 11.30, t(47) = 2.23, p = .003$.

**Modelling**

One of the key parts of this thesis is evaluating the ability of the EXIT model to capture IBRE data sets. As such, the EXIT model was used to simulate specific experiments within this thesis, aiming to capture the response patterns for each stimulus in each condition. The version of EXIT used for the simulations in this thesis is slpEXIT, a function contained within the *catlearn* R package (Wills et al., 2019). This implementation is based on the model as described in Kruschke (2001b), with the inclusion of a bias cue that was later implemented in Kruschke (2003). The model contains 7 free parameters as detailed within the description of EXIT in Chapter 1.

As the experiments within this thesis differ, the specific details of each experimental
procedure were captured through the generation of simulated training and test trials for each experiment. The EXIT model was then applied to these simulated training and test trials generating response patterns for each simulated trial. The values of the free parameters given to the model were varied using the optim function from the stats package in R (R Core Team, 2018). The goal of this variation was to optimise the free parameters given to the model for each experiment; in order to find the parameter set that when given to the model gave the closest approximation to the each experiment's behavioural data. This was accomplished by calculating the sum of squared errors (SSE) between the response patterns generated by the model under a specific parameter set and the behavioural response patterns. optim was used to find the parameter set that minimised the SSE.

Wills and Pothos (2012) make several cogent points about the advantages of finding a common set of parameters for a model that fits data from multiple experiments. There were two primary reasons for not doing this; first these are novel data sets and so this is a first-pass test of the EXIT model on this data. In this case failure to capture results using individually-fit parameters for each experiment still represents a failure of the model. Secondly, these model fits are not looking to evaluate model adequacy compared to other formally specified models. Rather these model fits are simply looking to evaluate if EXIT, a model specifically designed to capture the IBRE, can capture these novel IBRE datasets when given the maximum flexibility in the variation of its free parameters to do so.

The method used for optimisation within optim in all simulations in this thesis is the limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Byrd, Lu, Nocedal, & Zhu, 1995), extended to handle bound constraints on parameters. As optim requires an initial set of parameters to vary, each free parameter within the EXIT model was initially set to one of two values. As there are 7 free parameters, this resulted in a total of $2^7$ or 128 sets of parameter values. Each of these starting parameter sets were supplied to optim individually. optim then used the BFGS algorithm to perform a hill-climbing optimisation and arrive at an optimised parameter set for each individual starting set. This produced 128 sets of optimised parameter values. These sets of optimised parameters values were compared in terms of the SSE generated when they were given to EXIT, in order to identify the set that produced the lowest SSE. The parameter values within this final optimised set are then reported in this thesis.
For the experiments within this chapter, the bias cue was assumed to be analogous to the experimental context. As the experiments have two contexts, two bias cues were used to represent this. Table 2.3 shows the optimal parameter values, as well as the Root Mean Squared Error (RMSE) and $R^2$ between the simulated and behavioural data when these parameters are given to the model. The optimal parameters for Experiment 1 suggest moderate reliance on the contextual elements of the experiment to make decisions; shown by the large value of the $\sigma$ parameter. The $c$ parameter, $P$ parameter and $\phi$ parameters are all large, suggesting that the exemplars within the model are not confusable with each other, that one cue attracts most of the attention and that the model is decisive in its responses respectively. The learning parameters, $\lambda_g$ and $\lambda_x$ are also large suggesting that the attentional shift and the learning rate for the weights between exemplars nodes within the model are both large. The value for the $\lambda_w$ parameter is moderate, indicating a moderate learning rate for the associative weights within the model.

The key response patterns that EXIT should capture are those for the BC, B, C and N cues. In this experiment, responses to BC are preferentially rare, rare responses to C are significantly greater than common responses to B, and responding to N switches from preferentially common to random responding shifting from the same to the different context conditions. Values within the brackets in Table 2.2 represent the simulated data produced by EXIT using the optimal parameters. EXIT captures the response patterns for the BC cue, the B and C cues and the N cue. Overall, the RMSE value is very low, indicating a good fit to the data and the $R^2$ is very high indicating that a large proportion of the variance is captured by the model.

**Discussion**

Experiment 1 found preferentially rare responding to BC in both same and different context conditions, showing an IBRE. There was no evidence of a B>C effect; in fact the ordinal trend of the data suggests a C>B effect in both context conditions. There was preferentially common responding to A in both conditions, with greater common
2.2. EXPERIMENT 1

responding in the different-context than the same-context condition. Finally, responding to N was at random. However, there was evidence of an effect of context shift, with greater common responding in the same-context condition than the different-context condition.

The lack of a difference between conditions for the IBRE and B>C effect does not support the predictions made by the context explanation. A potential criticism of this experiment is that the context shift was not salient enough to elicit a change in between conditions. However, the response pattern for the N cue effectively refutes this criticism. There is a significant effect of context shift for the N cue, following the change in response pattern that the context explanation predicted. The fact that only the predicted change of response pattern for the novel cue was found in this experiment is a major issue for the context explanation. Further, greater rare responding to C than common responding to B, as well as greater common responding to A in the different-context relative to the same-context condition is the opposite of what the context explanation predicts.

The EXIT model is able to capture the ordinal trends of the behavioural data for Experiment 1 well. It provides a close fit of all of the key results for the BC, B, C and N cues. This effectively demonstrates the strength of EXIT and the error-driven attentional account of the IBRE.

The ordinal response pattern for the novel cue is interesting in this experiment, given that previous work suggests preferentially rare responding rather than common (Johansen et al., 2007; Juslin et al., 2001). However, these contradictory findings are explainable considering the statistical evidence. Neither Juslin et al. (2001) or Johansen et al. (2007) report any statistical analysis for the novel cue and the preferentially common responding to the novel cue in Experiment 1 is not significant.

The findings of Experiment 1 suggest that the context explanation is not adequate. However, the procedure of Experiment 1 did not produce the IBRE and B>C effect in conjunction. It would be premature to dismiss the context explanation at this point; given that the lack of a difference could be due to the failure to find the IBRE and the B>C effect rather than a failure to demonstrate the reversal predicted by the context explanation. Due to this, Experiment 2 continues the evaluation of the context explanation.
2.3. EXPERIMENT 2

Experiment 2

There are several potential reasons for why Experiment 1 failed to find the IBRE and $B > C$ effect in conjunction. The task participants were required to complete was novel in terms of IBRE experimental procedures, and making use of criterion-based training is not something IBRE studies commonly do. The criterion in Experiment 1 was difficult; requiring near perfect accuracy in order to pass training. This potentially caused participants to be experts with both stimuli types, hence the lack of a $B > C$ effect as participants were familiar with both cues. To address these points, Experiment 2 was changed in two ways. First, the nature of the task was changed from diagnosing animals back to the more classical human diagnosis task. Second, instead of removing the criterion, it was decided to instead reduce the difficulty of passing it. This should determine if the issue with the criterion is the difficulty of it or the criterion itself. A context manipulation was still implemented within Experiment 2, opting to task participants with diagnosing humans during training (same-context) and either humans or orcs (different-context) at test. The key goals for this experiment remain the same as for Experiment 1; to produce the IBRE and $B > C$ effect in conjunction, to examine the extent to which the context explanation can explain the behavioural data and to assess the ability of the EXIT model to capture the behavioural data patterns.

Method

Participants

72 people participated in this experiment.

Stimuli

The abstract design of Experiment 2 was the same as Experiment 1 (see Table 2.1). The cues were again symptoms, but consisted of: ear aches, skin rash, back pain, dizziness, sore muscles, stuffy nose, nausea, fever. The symptom paired with the abstract cue was randomised for each participant. For the training phase, each of the cue compounds led to a different outcome; one of four diseases: F, G, H and J. The diseases were mapped to the abstract disease types in every possible way, across participants.

The test phases were completed under both a human context and an orc context. For the human context, the cues were presented in the centre of the screen with no obvious indications other than the instructions that they were diagnosing humans. For the orc
2.3. **EXPERIMENT 2**

![Diagram of an orc's head with symptoms and question: Which disease does the patient have? (F, G, H, J)]

Figure 2.4: Example of how the cues were presented during the orc context of Experiment 2

context, in order to emphasise the differing context, the cues were instead within a caricatured outline of an orc’s head. Figure 2.4 demonstrates how the cues were presented for the orc context.

**Procedure**

The experimental design of Experiment 2 was close to that of Experiment 1. Participants were still asked to diagnose patients, but the two possible contexts were switched from cats/dogs to humans/orcs. Further, participants were only trained on the human context rather than both. In this thesis, I refer to the human context as “X” due to similarity between this procedure and the allergist paradigm, which commonly employs a “Mr. X” (Aitken, Larkin, & Dickinson, 2000). The orc context is referred to as “O” for obvious reasons. As such, the two possible conditions participants could receive are labelled as “XXO” and “XOX”, based on the order of context experienced in the training and two test phases. In each of the phases listed below the trial order was randomised between participants.

The training phase instructions, abstract trial design, and stimulus timings were identical to Experiment 1. Participants completed a training block of 16 trials before being assessed according to a criterion. Participants were informed of what percentage of trials they had diagnosed correctly, and that they needed 90% correct to progress. If participants failed to meet this criterion they received another training block, up to a maximum of 15. Once a participant met this criterion they were moved onto the first test phase.
The instructions for the test phase were the same as Experiment 1, apart from the following exceptions. The instructions for those in the XOX condition emphasised the context change by telling participants they would be completing a medical placement in a different dimension. They also received an example trial before the test phase so that they could understand the trial structure. This example trial consisted of the orc head being presented for 10000 ms with the word “Symptoms” displayed in the centre of it, and the rest of the text present on the training trials in the same places as before. As in Experiment 1, feedback was unavailable. After making a response, they now received a screen telling them “No diagnostic feedback report available. Your response has been recorded.” for 2000 ms. Participants received 90 trials before being moved on to the second test phase.

Participants received a final set of instructions. In the XXO condition, they were informed that they would now be completing a medical placement in another dimension and would now be diagnosing orcs. They received the example orc trial before the test phase as the XOX participants did before their first test phase. XOX participants were informed that they had completed their placement and would now again be diagnosing humans. The trial structure was the same as that of the first test phase, and after participants completed another 90 trials the experiment finished.

Results

The trial-level raw data, analysis scripts and modelling scripts for this experiment will be available at https://osf.io/uy3qq/ once this experiment is published. Participants completed one of the two conditions, with 48 participants completing the XXO condition and 24 participants completing the XOX condition. Participants’ training phase performance was assessed in terms of the number of training blocks it took for them to pass the learning criterion. The mean number of blocks to pass criterion was 4.08 (SD = 2.61) and all participants did eventually reach criterion.

Looking at participants’ accuracy in the first and last blocks of training, it was found that accuracy in the final block was significantly higher than accuracy in the first block, $F(1,71) = 159.69, p < .001$, and participants were more accurate at the common stimuli than the rare stimuli, $F(1,71) = 19.23, p < .001$. No significant interaction between the

---

1Due to the difficulty of finding both IBRE and B>C in Experiment 1, twice the number of people were run in XXO compared to XOX. As there is no context shift in the first test phase this seems the purest condition (i.e. least likely to be affected by experimental context manipulation) and therefore most likely to find the IBRE and B>C effect in conjunction.

2It’s worth noting that in comparison to Experiment 1, the training blocks of Experiment 2 contained half the number of trials. Therefore 4.08 training blocks in Experiment 2 is 2.04 blocks in Experiment 1.
2.3. EXPERIMENT 2

Table 2.4: Proportion of responses to each of the stimulus types presented under different-context and same-context conditions for Experiment 2. Bold type highlights the results of primary theoretical interest. Values within brackets represent response proportions from the simulation of this experiment using the EXIT model with optimised parameters.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>Common</th>
<th>Rare</th>
<th>Common Other</th>
<th>Rare Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same</td>
<td>Diff</td>
<td>Same</td>
<td>Diff</td>
</tr>
<tr>
<td>A</td>
<td>.78(.69)</td>
<td>.74(.69)</td>
<td>.12(.13)</td>
<td>.17(.13)</td>
</tr>
<tr>
<td>AB</td>
<td>.90(.90)</td>
<td>.89(.90)</td>
<td>.05(.04)</td>
<td>.06(.04)</td>
</tr>
<tr>
<td>ABC</td>
<td>.56(.65)</td>
<td>.59(.65)</td>
<td>.36(.25)</td>
<td>.34(.25)</td>
</tr>
<tr>
<td>AC</td>
<td>.28(.27)</td>
<td>.31(.27)</td>
<td>.60(.59)</td>
<td>.58(.59)</td>
</tr>
<tr>
<td>B</td>
<td>.88(.82)</td>
<td>.86(.81)</td>
<td>.06(.05)</td>
<td>.07(.05)</td>
</tr>
<tr>
<td>BC</td>
<td>.41(.41)</td>
<td>.44(.41)</td>
<td>.49(.43)</td>
<td>.48(.44)</td>
</tr>
<tr>
<td>C</td>
<td>.10(.05)</td>
<td>.14(.05)</td>
<td>.73(.82)</td>
<td>.70(.82)</td>
</tr>
<tr>
<td>N</td>
<td>.42(.51)</td>
<td>.46(.50)</td>
<td>.58(.49)</td>
<td>.54(.50)</td>
</tr>
</tbody>
</table>

Factors was found, $F(1,71) = 1.76, p = .19$. Looking at the final block of training, participants were significantly more accurate at the *common* stimuli compared to the *rare* stimuli, $t(71) = 3.77, p < .001$.

Table 2.4 shows the response proportions for each stimulus under both different-context and same-context conditions. Due to increased *rare-other* response proportions for the C stimulus, the *common* and *rare* response types were not rescaled for the B and C comparison. Looking first at the same-context condition, *rare* responses to the BC stimulus did not differ from .5, $BF = .15, t(71) = 0.50, p = .62$, indicating the absence of the IBRE. *Common* responses to the B stimulus were significantly higher than *rare* responses to the C stimulus, $BF = 303.87, t(71) = 4.24, p < .001$, indicating the presence of the B>C effect.

The proportion of *common* responses to the A stimulus was significantly greater than .5, $t(71) = 16.61, p < .001$. *Common* responses to the novel N stimulus suggests that they did not differ from .5, $BF = .40, t(71) = 1.54, p = .13$, although the Bayes is inconclusive.

Looking second at the different-context condition, *rare* responses to the BC stimulus did not differ from .5, $BF = .13, t(71) = 0.18, p = .86$, indicating the absence of the IBRE. *Common* responses to the B stimulus were again significantly higher than *rare* responses to the C stimulus, $BF = 124.30, t(71) = 3.97, p < .001$, indicating the presence of the B>C effect. The proportion of *common* responses to the A stimulus was significantly greater than .5, $t(71) = 11.78, p < .001$. *Common* responses to the novel N stimulus did not differ from .5, $BF = .18, t(71) = 0.81, p = .42$.

Looking third at the difference in response proportions between conditions, there was no difference in the proportion of *rare* responses between conditions for BC stimuli,
2.3. EXPERIMENT 2

\[ BF = .15, t(71) = .59, p = .56. \] Similarly there was no difference between the conditions for common responses to the B stimulus, \[ BF = .17, t(71) = .74, p = .46. \] For the C stimulus there was also no difference between the conditions, \[ BF = .47, t(71) = 1.64, p = .11, \] but the Bayes evidence is inconclusive. There was significantly greater common responding in the same-context compared to the different-context condition for the A stimulus, \[ t(71) = 2.66, p = .009. \] There was no evidence of an effect of context shift on the novel N stimulus, \[ BF = 0.15, t(71) = 0.58, p = .56. \]

Modelling

Table 2.5 shows the optimal parameter values for Experiment 2, as well as the Root Mean Squared Error (RMSE) and \( R^2 \) between the simulated and behavioural data when these parameters are given to the model. The optimal parameters for Experiment 2 show a low value for the \( \sigma \) parameter suggesting that a high salience for context is not prioritised in the optimisation of this experiment. The \( c \) parameter, \( P \) parameter and \( \phi \) parameter values suggest that the exemplars are less confusable with each other, that one cue attracts most of the attention and that the model is decisive in its responses respectively. The learning parameters, \( \lambda_g, \lambda_w \) and \( \lambda_x \) suggest that the attentional shift within the model is large, that the rate at which associative weights are learnt is moderate and that the learning rate for the weights between exemplar nodes and gain nodes is large.

In this experiment, responses to BC are at random, common responses to B are significantly greater than rare responses to C, and responses to N are at random. Values within the brackets in Table 2.4 represent the simulated data produced by EXIT using the optimal parameters. EXIT captures the response patterns for the BC cue and the N cue. However, it is unable to capture common responses to B being significantly greater than rare responses to C. One further point about the behavioural data worth mentioning is that in this experiment there is evidence of an effect of context on the A cue. EXIT does not capture this effect as can be seen in Table 2.4. Overall, the RMSE value is very low, indicating a good fit to the data and the \( R^2 \) is very high indicating that a large proportion

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( c )</th>
<th>( P )</th>
<th>( \phi )</th>
<th>( \lambda_g )</th>
<th>( \lambda_w )</th>
<th>( \lambda_x )</th>
<th>( \sigma )</th>
<th>RMSE</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.353</td>
<td>2.087</td>
<td>3.367</td>
<td>3.683</td>
<td>0.108</td>
<td>2.238</td>
<td>0.138</td>
<td>0.046</td>
<td>0.973</td>
</tr>
</tbody>
</table>
of the variance is captured by the model.

**Discussion**

Experiment 2 found random responding to BC in both same and different context conditions, showing a lack of an IBRE in this experiment. *Common* responding to B was significantly greater than *rare* responding to C in both same and different context conditions. Responding to A was *common* in both conditions, with greater *common* responding found in the same-context condition relative to the different-context condition. Finally, responses to the N cue were random in both context conditions.

The context explanation struggles with these results as well as those of Experiment 1. It would assume that the learned association between context cue and the *common* outcome is present given the evidence of a B>C effect and the reduction of *common* responding to the A cue in the different-context condition relative to the same-context condition. However, this then doesn’t explain the random responding to the BC and N cues.

EXIT is again able to capture the majority of the behavioural response patterns in this experiment. In terms of the key results, it captures both the BC and N response patterns. However, one sticking point is that it doesn’t capture the B>C effect. This is a known issue with EXIT; addressed in Kruschke (2003). EXIT is able to capture the B>C effect if the response pattern is more heavily weighted in relation to the other effects the model fit aims to capture. This was not implemented for the simulations within this thesis, on the basis that it might adversely affect the model’s ability to capture novel cue response patterns. Another issue is that EXIT does not predict a difference in responding *common* to the A cue between conditions, something the behavioural data finds significantly. The model does however capture the general response pattern to the A cue noted in other IBRE studies.

Experiment 2 again fails to find both the IBRE and B>C effect in conjunction, under an experimental procedure closer to that of classical IBRE studies. Again, it would be premature to reject the context explanation; the failures of the explanation in this experiment could still be due to failing to find the IBRE and B>C effect. As a result of this, Experiment 3 takes a step back in terms of what it aims to accomplish; solely seeking to identify a procedure capable of producing the effects in conjunction.
Experiment 3

As previously mentioned one of the key differences from classical IBRE studies present in the procedures of Experiments 1 and 2 was criterion-based learning. Whilst Experiment 2 aimed to reduce the difficulty of this, Experiment 3 aims to implement a procedure that does not make use of it. At this point, the replicability of the B>C effect is also called into question. Given the low number of studies that demonstrate it (Bohil et al., 2005; Medin & Edelson, 1988; Wills et al., 2014; Winman et al., 2005), as well as the procedural differences between these studies, it is possible that the effect is unreliable.

Experiment 3 addresses both these problems by directly replicating the procedure of the most recent study to demonstrate the B>C effect; Wills et al. (2014). This study used an ERP methodology to investigate brain potentials arising as a result of the IBRE. Rather than introduce the complexity of an ERP methodology, Experiment 3 uses the same behavioural procedure of Wills et al. (2014). The procedure has a fixed training length; addressing the potential issue of criterion-based learning. Participants are tasked with diagnosing patients based on presented “cells”, abstract coloured shapes, instead of symptoms. No context manipulation is implemented here to make the procedure as close to the original as possible, and no novel cue is present either due to it not being present in the original study.

The key goal of this experiment is singular; to produce the IBRE and B>C effect in conjunction. If this experiment cannot find these effects together then the B>C effect is not reliable enough to be used as a metric for evaluating the context explanation. It would also call into question the need for the explanation as a whole. One further point; no simulation of EXIT was conducted for the results of this experiment. As previously mentioned, EXIT has captured the IBRE and B>C effect before (Kruschke, 2003), and so simulating this experiment would be redundant.

Method

Participants

24 people participated in this experiment.

Stimuli

The stimuli used in this experiment were presented against a black background and were the same as those used in Wills et al. (2014).
Table 2.6: Abstract trial types for the training and test phases of Experiment 3. Bold type highlights the test stimuli of primary theoretical interest. The training trials appear in the test phase in the same proportion and participants continue to receive feedback at test for these trials to maintain learning.

<table>
<thead>
<tr>
<th>Training trials (relative frequency)</th>
<th>Test trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1B_1 \rightarrow$ common (x 2)</td>
<td>$A_1B_1, A_2B_2, A_3B_3$, $F_1D_1, F_2D_2, F_3D_3$, $x \ 2$</td>
</tr>
<tr>
<td>$A_2B_2 \rightarrow$ common (x 2)</td>
<td>$B_1, B_2, B_3, C_1, C_2, C_3$, $D_1, D_2, D_3, E_1, E_2, E_3$</td>
</tr>
<tr>
<td>$A_3B_3 \rightarrow$ common (x 2)</td>
<td>$A_1C_1, A_2C_2, A_3C_3$, $G_1E_1, G_2E_2, G_3E_3$, $x \ 1$</td>
</tr>
<tr>
<td>$A_1C_1 \rightarrow$ rare (x 1)</td>
<td>$A_1C_1, A_2C_2, A_3C_3$, $G_1E_1, G_2E_2, G_3E_3$, $x \ 1$</td>
</tr>
<tr>
<td>$A_2C_2 \rightarrow$ rare (x 1)</td>
<td>$A_1C_1, A_2C_2, A_3C_3$, $G_1E_1, G_2E_2, G_3E_3$, $x \ 1$</td>
</tr>
<tr>
<td>$A_3C_3 \rightarrow$ rare (x 1)</td>
<td>$A_1C_1, A_2C_2, A_3C_3$, $G_1E_1, G_2E_2, G_3E_3$, $x \ 1$</td>
</tr>
<tr>
<td>$F_1D_1 \rightarrow$ common (x 2)</td>
<td>$B_1C_1, B_2C_2, B_3C_3$, $x \ 1$</td>
</tr>
<tr>
<td>$F_2D_2 \rightarrow$ common (x 2)</td>
<td>$D_1E_1, D_2E_2, D_3E_3$, $x \ 1$</td>
</tr>
<tr>
<td>$F_3D_3 \rightarrow$ common (x 2)</td>
<td>$A_1, A_2, A_3$</td>
</tr>
</tbody>
</table>

This led to the exclusion of the ABC trial types used in the previous two experiments, given that Wills et al. (2014) did not use them. Another significant change from Experiments 1 and 2 is the inclusion of disjointed cue trials, FD and GE, where the proportion of outcomes is the same as AB and AC but there is no shared cue. Table 2.6 shows the combinations of abstract cues and diseases presented in the training and test phases of the experiment. The cues themselves are a number of “cells”, red and yellow in colour and 30 x 30 pixels in size. An example of the “cells” on an experimental trial can be seen in Figure 2.5.

The stimuli were assigned at random to one of 7 abstract cues (A-G) for each participant.

Figure 2.5: An example trial presentation from the training phase of Experiment 3
As in the Wills et al. (2014) paper, each abstract cue had 3 stimuli assigned to it resulting in 21 different “cells” being used. A subset of possible cell combinations was used for the compound cue trials, for example on AB trials the cells presented were: $A_1 B_1$, $A_2 B_2$ or $A_3 B_3$. Approximate stimulus size was 2 degrees of visual angle for the “cells”. Two diseases were used as outcomes in this experiment: “Jominy Fever” and “Phipps Syndrome”. Diseases were mapped to the abstract disease types in both possible ways, across participants.

**Procedure**

In each of the phases listed below the trial order was randomised between participants. The instructions for Experiment 3 were similar to those of the previous experiment. Significant differences include no contextual manipulation, completing the task on the basis of “cells” rather than symptoms, responding on each trial with either “Jominy Fever” or “Phipps Syndrome” and the use of “c” and “m” as response keys. Participants were informed of the disease represented by each response key through their consent forms. The keys that represented “Jominy Fever” and “Phipps Syndrome” were fully counterbalanced between participants.

Participants first completed a training phase consisting of 20 blocks of 18 trials, resulting in 360 trials in total. Figure 2.6 shows the structure of a trial in Experiment 3. Each trial began with a 1000 ms presentation of a grey view box, intended to illustrate where the “cells” would be appearing. After this the “cells” for that trial appeared in the view box, centralised horizontally, but towards the top and bottom of the view box vertically. The “cell” presented to the top and to the bottom of the view box was randomised on each trial. The “cells” remained on screen for a maximum of 2000 ms, during which time participants

![Figure 2.6: The structure of a trial in Experiment 3. The view box is displayed from 0 - 3000ms. The “cell(s)” is/are displayed from 1000 - 3000ms. The feedback is displayed from 3000 - 4500ms. If a response was made while the “cell(s)” and view box were on screen, then the “cell(s)” and view box disappeared, and feedback was displayed for 1500ms from that point.](49)
made their diagnosis using the “c” and “m” keys. Once a response was made, the “cells” disappeared and participants received corrective feedback, telling them if they were right or wrong and what the correct diagnosis was. Feedback was presented on screen for 1500 ms before a new trial began. If a response was not made during the 2000 ms the “cells” were on screen, then participants instead received a time-out message telling them they were out of time and to speed up. The time-out message was displayed for the same duration as the feedback.

After completing the training phase, participants received a second set of instructions. Again these instructions were similar to those previously used in Experiments 1 and 2; but instead of receiving no feedback, participants received feedback on those trials they had seen during training. Participants then completed a test phase of 8 blocks of 51 trials; 408 trials in total. The trial structure in the test phase was the same as in the training phase, but with single “cells” being presented in the centre of the view box. On trials for which participants did not receive feedback, they received the message “DATA MISSING” and a series of question marks.

Results

The trial-level raw data and analysis scripts for this experiment will be available at https://osf.io/z5pqn/ once this experiment is published. Wills et al. (2014) assessed participants’ training accuracy according to a criterion, where participants scoring less than 72% correct on the final block of training were removed from analysis. This 72% correct criterion represents the level of accuracy that cannot be attributed to random responding based on the block length of 18 trials. Applying this criterion here necessitated the removal of 2 participants, resulting in a final data set of 22 participants.

The accuracy of participants across the training phase can be seen in Figure 2.7. Looking at the accuracy in the first and final blocks of training, it was found that accuracy in the final block was significantly higher than accuracy in the first block, $F(1,21) = 104.81, p < .001$. No significant effects were found for stimulus frequency, $F(1,21) = 2.19, p = .15$, or shared cue, $F(1,21) = 3.25, p = .09$. A significant interaction between block and shared cue was found, $F(1,21) = 5.27, p = .03$, indicating that accuracy was greater in the final block and that this difference in accuracy was greater for non-shared cue stimuli. No significant interactions were found for any combination of the other factors. Looking at the final block of training alone, accuracy was significantly higher for the common stimuli compared to
Figure 2.7: Participants’ accuracy on the abstract training trial types at different levels of training in Experiment 3. The error bars represent within-subject Cousineau-Morey 95% confidence intervals.

The rare stimuli, $F(1, 21) = 7.10, p = .01$, and on the non-shared cue stimuli over the shared cue stimuli, $F(1, 21) = 7.49, p = .01$. No significant interaction between these two factors was found.

Table 2.7 shows the response proportions for each of the stimuli presented in the test phase. Looking at the BC stimulus, NHST suggests that rare responses are significantly greater than .5, $BF = 1.59, t(21) = 2.19, p = .04$. This indicates the presence of an IBRE, although the Bayes is inconclusive. When testing the proportion of common responses to the B

Table 2.7: Proportion of responses to each of the stimulus types presented in the test phase of Experiment 3. Bold type highlights the results of primary theoretical interest.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.77</td>
<td>.23</td>
</tr>
<tr>
<td>AB</td>
<td>.90</td>
<td>.10</td>
</tr>
<tr>
<td>AC</td>
<td>.26</td>
<td>.74</td>
</tr>
<tr>
<td>B</td>
<td>.92</td>
<td>.08</td>
</tr>
<tr>
<td>BC</td>
<td>.39</td>
<td>.61</td>
</tr>
<tr>
<td>C</td>
<td>.23</td>
<td>.77</td>
</tr>
<tr>
<td>D</td>
<td>.90</td>
<td>.10</td>
</tr>
<tr>
<td>DE</td>
<td>.42</td>
<td>.58</td>
</tr>
<tr>
<td>E</td>
<td>.27</td>
<td>.73</td>
</tr>
<tr>
<td>FD</td>
<td>.96</td>
<td>.04</td>
</tr>
<tr>
<td>GE</td>
<td>.14</td>
<td>.86</td>
</tr>
</tbody>
</table>
stimulus against rare responses to the C stimulus, the proportion of common responses was found to be significantly higher, BF = 5.76, t(21) = 2.91, p = .01, indicating the presence of the B>C effect. Similarly, the proportion of common responses to the A stimulus was found to be significantly greater than .5, t(21) = 5.16, p < .001. The DE stimulus (shared cue test) was not found to be significantly different from .5, t(21) = 1.35, p = .19. The proportion of common responses to the BC stimulus was not found to be significantly different from the proportion of common responses to the DE stimulus, t(21) = .71, p = .49.

Discussion

Experiment 3 found random responding to the BC test cue, failing to demonstrate an IBRE. Common responding to B was significantly greater than rare responding to C, showing the presence of the B>C effect. Therefore this experiment does not meet its key outcome; it does not find a procedure that produces the IBRE and B>C effect in conjunction.

Another response pattern noted in Wills et al. (2014) that does not replicate is that of the DE stimulus. When presented at test, it shows random responding in this experiment; whilst responding was preferentially common in the original study. During the process of developing this experiment, a possible procedural error was discovered through re-examination of the file archive in our lab. Specifically, stimuli labelled as DE within the files appear instead to have been presented as FD stimuli. FD stimuli were presented during training with feedback and led to the common outcome. If this procedural error was indeed present then it would explain the preferentially common responding observed in Wills et al. (2014).

Despite the lack of a significant IBRE, the response patterns for the BC, B and C cues do suggest both the IBRE and B>C effect ordinally. Considering the statistical analysis for BC, NHST testing does suggest significant rare responding to the BC cue; it is the Bayes that is inconclusive, suggesting responding is at random. As both the present experiment and Wills et al. (2014) have low numbers of participants, it is perhaps unsurprising that the effect is found significantly in Wills et al. (2014) and not in the present experiment. Of the three experiments reported in this chapter so far, the results of Experiment 3 are the most consistent with previous work finding both the IBRE and B>C effect (Bohil et al., 2005; Wills et al., 2014; Winman et al., 2005). As such, Experiment 4 builds on Experiment 3; increasing the number of participants and adding a context manipulation to the procedure to finally fully evaluate the context explanation.
Experiment 4

The procedure for Experiment 4 combines the procedures used in Experiments 2 and 3. Given that the procedure of Experiment 3 is the best candidate for producing the IBRE and B>C effect in conjunction, the basic design is that of Experiment 3. The context manipulation is adapted from Experiment 2, diagnosing humans at training and both humans and orcs in separate test sessions.

The key goals for Experiment 4 are as follows: produce the IBRE and B>C effect in conjunction, evaluate the context explanation in terms of its ability to account for the behavioural data and evaluate the ability of EXIT to capture the behavioural response patterns within this experiment.

Method

Participants

94 people participated in this experiment.

Stimuli

The stimuli used can be seen in Table 2.8 and were the same as in Experiment 3. The exception to this is the addition of the novel cue, N. The N cue has six stimuli assigned to it, rather than the three assigned to the other abstract cues. This is due to the novel nature of the cue and the fact that there are two test phases;

<table>
<thead>
<tr>
<th>Training trials (relative frequency)</th>
<th>Test trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1B_1 \rightarrow \text{common (x 2)}$</td>
<td>$A_1B_1, A_2B_2, A_3B_3, \text{ x 2}$</td>
</tr>
<tr>
<td>$A_2B_2 \rightarrow \text{common (x 2)}$</td>
<td>$F_1D_1, F_2D_2, F_3D_3, B_1, B_2, B_3, C_1, C_2, C_3, \text{ x 1}$</td>
</tr>
<tr>
<td>$A_3B_3 \rightarrow \text{common (x 2)}$</td>
<td>$D_1, D_2, D_3, E_1, E_2, E_3$</td>
</tr>
<tr>
<td>$A_1C_1 \rightarrow \text{rare (x 1)}$</td>
<td>$A_1C_1, A_2C_2, A_3C_3, \text{ x 1}$</td>
</tr>
<tr>
<td>$A_2C_2 \rightarrow \text{rare (x 1)}$</td>
<td>$G_1E_1, G_2E_2, G_3E_3, B_1C_1, B_2C_2, B_3C_3, \text{ x 1}$</td>
</tr>
<tr>
<td>$A_3C_3 \rightarrow \text{rare (x 1)}$</td>
<td>$D_1E_1, D_2E_2, D_3E_3, A_1, A_2, A_3, \text{ x 1}$</td>
</tr>
<tr>
<td>$F_1D_1 \rightarrow \text{common (x 2)}$</td>
<td>$N_{1</td>
</tr>
<tr>
<td>$F_2D_2 \rightarrow \text{common (x 2)}$</td>
<td>$N_{1</td>
</tr>
<tr>
<td>$F_3D_3 \rightarrow \text{common (x 2)}$</td>
<td>$N_{1</td>
</tr>
<tr>
<td>$G_1E_1 \rightarrow \text{rare (x 1)}$</td>
<td>$N_{1</td>
</tr>
<tr>
<td>$G_2E_2 \rightarrow \text{rare (x 1)}$</td>
<td>$N_{1</td>
</tr>
<tr>
<td>$G_3E_3 \rightarrow \text{rare (x 1)}$</td>
<td>$N_{1</td>
</tr>
</tbody>
</table>
resulting in three of the cues assigned to N being used in the first test phase and the other three in the second. This resulted in 27 different “cells” being used in total.

Procedure

Experiment 4 is an extension of the procedure used in Experiment 3, accommodating the contextual elements of Experiment 2. The contexts used are humans and orcs, with participants only being trained in the human context. As in Experiment 2, the two conditions are labelled XXO and XOX based on the order of context experienced in the training and two test phases. The instructions that participants received at the start of the experiment are the same as those given in Experiment 3. In each of the phases listed below the trial order was randomised between participants.

Figure 2.8 shows the structure of a trial in Experiment 4. The training phase was identical to Experiment 3. After completing training, participants received a second set of instructions. In the XXO condition, the instructions were the same as the test phase of Experiment 3. In the XOX condition participants were informed that they would be diagnosing orcs, in the same manner as the instructions in the XOX condition of Experiment 2. These participants received an example trial with an orc patient, with the orc head used in Experiment 2 surrounding the view box that appears on each trial.

Each test phase consisted of 216 trials. After completing the first test phase, participants received another set of instructions. These instructions were the ones presented to participants in the other condition for the first test phase. Those now diagnosing orcs also received an example trial as detailed above.
Figure 2.9: Participants’ accuracy on the abstract training trial types at different levels of training in Experiment 4. The error bars represent within-subject Cousineau-Morey 95% confidence intervals

Results

The trial-level raw data, analysis scripts and modelling scripts for this experiment will be available at https://osf.io/8p42b/ once this experiment is published. Participants completed one of the two conditions, with 48 participants completing the XXO condition and 46 participants completing the XOX condition. The same criterion as in Experiment 3 was applied, resulting in 42 participants left in the “XXO” condition and 34 participants left in the “XOX” condition.

The accuracy of participants across the training phase can be seen in Figure 2.9. Looking at accuracy in the first and last blocks of training, it was found that accuracy in the final block was significantly higher than accuracy in the first block, $F(1,75) = 312.90, p < .001$. Participants were significantly more accurate with the common stimuli than the rare stimuli, $F(1,75) = 9.78, p = .003$, and with the non-shared cue than the shared cue, $F(1,75) = 18.40, p < .001$. A significant interaction was found between frequency and shared cue, $F(1,75) = 6.85, p = .01$, indicating that accuracy was greater for non-shared cue stimuli and that this difference in accuracy was greater for rare outcome stimuli. Another significant interaction was found between frequency and block, $F(1,75) = 13.37, p < .001$, indicating that accuracy was greater in the final block and that this difference in accuracy was greater.

A Chi-square test of independence was calculated on the exclusion rate between conditions, considering the exclusion rate was double in the “XOX” condition. No significant interaction was found, $BF = .78, \chi^2(1) = 2.80, p = .09$. Similarly no significant interaction was found when comparing the overall exclusion rate between Experiment 3 and Experiment 4, $BF = .32, \chi^2(1) = .95, p = .37$.
Table 2.9: Proportion of responses to each of the stimulus types presented under different-context and same-context conditions for Experiment 4. Bold type highlights the results of primary theoretical interest. Values within brackets represent response proportions from the simulation of this experiment using the EXIT model with optimised parameters.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same</td>
<td>Diff</td>
</tr>
<tr>
<td>A</td>
<td>.72(.69)</td>
<td>.69(.61)</td>
</tr>
<tr>
<td>AB</td>
<td>.89(.91)</td>
<td>.87(.77)</td>
</tr>
<tr>
<td>AC</td>
<td>.30(.09)</td>
<td>.30(.20)</td>
</tr>
<tr>
<td>B</td>
<td>.87(.89)</td>
<td>.88(.83)</td>
</tr>
<tr>
<td>BC</td>
<td>.42(.40)</td>
<td>.42(.46)</td>
</tr>
<tr>
<td>C</td>
<td>.26(.16)</td>
<td>.26(.25)</td>
</tr>
<tr>
<td>D</td>
<td>.90(.86)</td>
<td>.88(.78)</td>
</tr>
<tr>
<td>DE</td>
<td>.47(.50)</td>
<td>.47(.51)</td>
</tr>
<tr>
<td>E</td>
<td>.25(.26)</td>
<td>.26(.36)</td>
</tr>
<tr>
<td>FD</td>
<td>.94(.91)</td>
<td>.92(.83)</td>
</tr>
<tr>
<td>GE</td>
<td>.14(.09)</td>
<td>.15(.17)</td>
</tr>
<tr>
<td>N</td>
<td>.57(.47)</td>
<td>.50(.51)</td>
</tr>
</tbody>
</table>

for common outcome stimuli. No other significant main effects or interactions between any combination of the other factors were found.

Looking at the accuracy in the final block of training alone found that participants were significantly better with the common stimuli than the rare stimuli, $F(1, 75) = 51.02, p < .001$, and at the non-shared cue than the shared cue, $F(1, 75) = 26.72, p < .001$. The interaction between frequency and shared cue was also significant, $F(1, 75) = 14.39, p < .001$, again indicating that accuracy was greater for non-shared cue stimuli and that this difference in accuracy was greater for rare outcome stimuli. No other significant main effects or interactions between any combination of the factors were found.

Table 2.9 shows the response proportions for each of the stimuli presented in the same- and different-context conditions. Looking first at the same-context condition, rare responses to the BC stimulus were found to be significantly higher than .5, $BF = 4.85, t(75) = 2.82, p = .006$, suggesting the presence of an IBRE. A t-test between the common response proportion to the B stimulus and the rare response proportion to the C stimulus found the common response to B to be significantly higher, $BF = 2709.15, t(75) = 4.87, p < .001$, showing the B>C effect. The proportion of common responses to the A stimulus was also significantly greater than .5, $t(75) = 9.32, p < .001$. The common responses to the DE stimulus were not significantly different from .5, $t(75) = 1.15, p = .26$. The proportion of common responses to the BC stimulus was not found to be significantly different from
the proportion of common responses to the DE stimulus, \( t(75) = 1.45, p = .15 \). NHST suggests that common responses to the N stimulus were significantly greater than .5, \( BF = 1.02, t(75) = 2.11, p = .038 \); although the Bayes is inconclusive.

Looking second at the different-context condition, rare responses to the BC stimulus were also found to be significantly higher than .5, \( BF = 14.15, t(75) = 3.23, p = .002 \), again suggesting the presence of an IBRE. A t-test between the common response proportion to the B stimulus and the rare response proportion to the C stimulus found the common response to B to be significantly higher, \( BF = 11111.51, t(75) = 5.25, p < .001 \), showing the B>C effect. The proportion of common responses to the A stimulus was again significantly greater than .5, \( t(75) = 8.30, p < .001 \). The common responses to the DE stimulus were not significantly different from .5, \( t(75) = 1.18, p = .24 \). The proportion of common responses to the BC stimulus was not found to be significantly different from the proportion of common responses to the DE stimulus, \( t(75) = 1.45, p = .15 \). Common responses to the N stimulus did not differ from .5, \( BF = .13, t(75) = .09, p = .92 \).

Looking third at the difference in response proportions between conditions, there was no difference in the proportion of rare responses between conditions for BC stimuli, \( BF = .13, t(75) = .31, p = .76 \). Similarly there was no difference between the conditions for common responses to the B stimulus, \( BF = .21, t(75) = 1.02, p = .31 \), or rare responses to the C stimulus, \( BF = .13, t(75) = .20, p = .84 \). There was also no difference between conditions for common responses to the A stimulus, \( t(75) = 1.26, p = .21 \), or DE stimulus, \( t(75) = .06, p = .95 \). There was evidence of an effect of context shift on the N stimulus, with greater common responding in the same-context condition than the different context condition, \( BF = 1.29, t(75) = 2.23, p = .03 \), although the Bayes is inconclusive.

**Modelling**

Table 2.10 shows the optimal parameter values, as well as the Root Mean Squared Error (RMSE) and \( R^2 \) between the simulated and behavioural data when these parameters are given to the model. The optimal parameters for Experiment 4 suggest that the model is making use of the context cue in generating its response patterns, as seen by a medium-high value for \( \sigma \). The \( c \) parameter suggests that the exemplars are slightly confusable with each other, whilst the P and \( \phi \) parameters show that one cue within the model attracts most of the attention on each trial and that the model is relatively decisive in its responses respectively. The learning parameters, \( \lambda_g, \lambda_w \) and \( \lambda_x \) suggest that the attentional shift within
2.5. EXPERIMENT 4

Table 2.10: Optimal parameters for the EXIT model for Experiment 4. RMSE and $R^2$ are calculated between simulated data produced by the model supplied with these parameters and the behavioural data.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$c$</th>
<th>$P$</th>
<th>$\phi$</th>
<th>$\lambda_m$</th>
<th>$\lambda_r$</th>
<th>$\lambda_x$</th>
<th>$\sigma$</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.980</td>
<td>3.788</td>
<td>2.313</td>
<td>1.065</td>
<td>0.406</td>
<td>2.912</td>
<td>0.655</td>
<td>0.075</td>
<td>0.931</td>
<td></td>
</tr>
</tbody>
</table>

the model is low, that the rate at which associative weights are learnt is large and that the learning rate for the weights between exemplar nodes and gain nodes is large.

In this experiment, responses to BC are preferentially rare, common responses to B are significantly greater than rare responses to C, and responding to N switches from preferentially common to random responding shifting from the same- to the different-context conditions. Values within the brackets in Table 2.9 represent the simulated data produced by EXIT using the optimal parameters. EXIT manages to capture the response patterns for the BC cue, the B and C cues and the novel cue N. Overall, the RMSE value is very low, indicating a good fit to the data and the $R^2$ is very high indicating that a large proportion of the variance is captured by the model. It is worth noting though that these values are higher and lower respectively compared to the previous experiment in this chapter.

Discussion

Experiment 4 found preferentially rare responding to BC in both same- and different-context conditions, showing an IBRE. Common responding to B was significantly greater than rare responding to C in both context conditions, again showing the B $>$ C effect. Responding to A was preferentially common in both conditions, with no evidence for an effect of context shift. Responses to N were ordinally more common in the same-context condition but random in the different-context condition. There is inconclusive evidence of an effect of context shift for the N cue between conditions.

The context explanation continues to fail to explain the behavioural data in this series of experiments. The only result it predicts is the response pattern for the novel cue being preferentially common, at least in the same context condition. However, unlike Experiment 1, it does not show a significant effect of context shift. The lack of a difference between context conditions for the BC, A, B and C cues again shows the inadequacy of context as an explanation of the IBRE and B $>$ C effect.
EXIT continues to demonstrate its strength, capturing both BC and N response patterns. It does fail to capture the B>C effect; but as previously stated this is a known issue.

As noted in Experiment 3, the DE stimulus again does not replicate the result reported in Wills et al. (2014). Given the results of the experiments within this chapter, it seems like preferentially common responding to DE is not a replicable aspect of Wills et al. (2014). The original result may have been due to technical error, as previously discussed.

Chapter Discussion

The experiments within this chapter aimed to do three things; to find a procedure capable of producing the IBRE and B>C effect in conjunction, to evaluate the predictions made by the context explanation in terms of their similarity to the behavioural data, and to test EXIT’s ability to capture behavioural response patterns for IBRE experiments; in particular the response pattern for the novel cue. Experiment 4 demonstrates a procedure capable of producing both effects. Overall the context explanation is inadequate as an explanation of the IBRE and B>C effect occurring in conjunction. Whilst it was unable to capture B>C response patterns, EXIT was generally able to capture the other ordinal patterns of the data within these experiments. Capturing B>C may indeed be solvable in the same manner as the approach detailed in Kruschke (2003). EXIT’s failure to capture an effect of context shift for the A cue in Experiment 2 may also be solvable in a similar manner; but resolving these issues is left for future work.

Whilst the predictions of the context explanation have been shown to be generally unsupported, it is interesting that its prediction for the novel cue response patterns is partially correct in Experiments 1 and 4. A potential explanation of this is based on the unique nature of the novel cue. Participants possess some knowledge about the outcome linked to every cue presented at test; with the exception of the novel cue. Given that participants have no previous experience with the novel cue they could potentially be recruiting known information about the experimental context in order to make a response for trials with a novel cue alone.

The contrast between the response patterns for the novel cue within the experiments in this chapter and those noted in Juslin et al. (2001) and Johansen et al. (2007) is still perplexing. One potential explanation for this considers the difference in training length between these experiments and the previous work. Participants in the experiments in this chapter experienced much more training than those in the studies of Juslin et al.
(2001) and Johansen et al. (2007). Therefore, it could be the case that the response pattern for the novel cue shifts from preferentially rare responding at low levels of training to preferentially common responding at higher levels of training. The pattern of results in the Juslin et al. (2001) paper partially supports this hypothesis, as the proportion of rare responding to the novel cue decreases with extra training. Nevertheless, further investigation is warranted.
Chapter 3

The effect of training length on the IBRE and a novel cue

Introduction

Whilst the length of training participants complete in IBRE studies does vary, Juslin et al. (2001) is the only study to directly manipulate training length and examine the effect this had on the IBRE. Their participants completed a training phase followed by a test phase, as is standard in IBRE procedures. This was then followed by a second training phase and a second test phase. Juslin et al. (2001) found an IBRE ordinarily in the first test phase, with responding to the BC cue being preferentially rare. However, the proportion of rare responses to the BC cue was reduced in the second test phase relative to the first, suggesting that the IBRE reduces as a result of further training. Juslin et al. (2001) also implemented a novel cue at test, finding that responding to it mirrored that of responding to the BC test cue. Responses were preferentially rare in the first test phase, with the proportion of these responses reducing in the second test phase relative to the first.

The issue with the results reported in Juslin et al. (2001) is simple; they do not report any statistical analysis of these results. This is compounded by the fact that no study has further examined training manipulation in an IBRE procedure, and few IBRE studies have employed a novel cue. Johansen et al. (2007) also find an ordinal pattern of preferentially rare responding to a novel cue at test, with the proportion of rare responding being much greater than in Juslin et al. (2001). However, no statistical analysis is reported for the novel cue in this study. Don and Livesey (2017) do employ a novel cue at test; but present it in compound with a shared cue present during training which is associated more heavily with the common outcome. This then makes the preferentially common responding to this compound at test unsurprising. Given the lack of studies and proper analysis investigating the effect of training on the IBRE and responding to a novel cue, it would be prudent to
more thoroughly study it before accepting these results as canonical.

At the end of Chapter 2 a potential reconciliation of the conflicting conclusions of Experiments 1 and 4, where responding to a novel cue was preferentially common, and Juslin et al. (2001) and Johansen et al. (2007), where responding to a novel cue was preferentially rare, was offered. This reconciliation raises the point that the length of training participants experienced in Chapter 2 was much greater than in either of the published studies. It is possible that both conclusions are correct, and that they occur after differing amounts of training.

The EXIT model (Kruschke, 2001a) is again fit to the experiments within this chapter. EXIT was able to fit the results of Chapter 2 well, including preferentially common responding to a novel cue. However, upon consideration of the architecture of the model, there should be no way for EXIT to produce preferentially rare responding to a novel cue. This is primarily because the only information EXIT would be able to use when presented with a novel cue is the association learned by the bias cue present on every trial, which due to the nature of the differing outcome frequencies in IBRE experiments, should be associated with the common outcome. As such, if the experiments within this chapter are able to demonstrate significant preferentially rare responding to a novel cue, then assessing EXIT’s ability to capture the results would be particularly interesting.

The aims of this chapter are twofold: one; to assess the effect of training on the IBRE and responding to a novel cue, and so the validity of the results reported in Juslin et al. (2001), and two; to evaluate if EXIT is able to capture the behavioural data of these experiments, particularly the response patterns to the novel cue.

**Experiment 5**

Experiment 5 is a first attempt to replicate the results found within Juslin et al. (2001). Much like the previous work, the procedure is within-subjects, with participants experiencing interleaved training and test blocks. Four levels of training were selected to capture the progression of the effect of training length from low levels of training through to a training level similar to that of Experiment 1 of Chapter 2.

Whilst the key cues in this experiment are the BC test cue and the novel cue, N, two more relevant results are also examined. The $B > C$ effect noted in Chapter 2 is also assessed here, as the results of Juslin et al. (2001) suggest that this effect would occur at low levels
of training but that the effect would disappear at higher levels of training with the proportion of common responding to B equalling the proportion of rare responding to C. A blank cue, K, is also implemented within this experiment. The purpose of this cue is to assess whether or not participants are using the base rates of the outcomes. The way in which this cue differs from the novel cue is that participants do not receive any information; yet they are expected to still give a response. If they make use of the base rates then they should respond preferentially common.

The key goals for this experiment are to replicate the reduction in the IBRE and rare responding to a novel cue noted in Juslin et al. (2001) and to assess EXIT’s ability to capture the behavioural data patterns of this experiment.

Method

Participants

46 people participated in this experiment.

Stimuli

The stimuli used were the same as in Experiment 1, with the exception of the addition of a blank cue, K, to give an indication of base responding rate. A blank cue is a trial on which no symptoms are presented. Table 3.1 shows the combinations of abstract cues and diseases presented in the training and test phases of the experiment. The novel cue, N, has four symptoms assigned to it, as there are four test phases and the cue must remain novel in each. The symptoms used were: chesty cough, fur loss, weepy eyes, vomiting, fatigue, diarrhoea, bad breath or skin ulcers. The symptom paired with each abstract cue was randomised for each participant. The four fictional disease names used were: Fips Syndrome, Gertz Syndrome, Haust Syndrome and Jominy Syndrome. Diseases were mapped to the abstract disease types in every possible way between participants.

Symptoms and diseases were presented in the context of either diagnosing a cat or

<table>
<thead>
<tr>
<th>Training trials (relative frequency)</th>
<th>Test trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1B_1 \rightarrow \text{common}_1$ (x 3)</td>
<td>$A_1,A_2,B_1,B_2,C_1,C_2$</td>
</tr>
<tr>
<td>$A_1C_1 \rightarrow \text{rare}_1$ (x 1)</td>
<td>$AB_1,AB_2,AC_1,AC_2$</td>
</tr>
<tr>
<td>$A_2B_2 \rightarrow \text{common}_2$ (x 3)</td>
<td>$BC_1,BC_2,ABC_1,ABC_2$, x 6</td>
</tr>
<tr>
<td>$A_2C_2 \rightarrow \text{rare}_2$ (x 1)</td>
<td>$N_1</td>
</tr>
</tbody>
</table>
3.2. EXPERIMENT 5

diagnosing a dog. To emphasise the context, the symptoms were presented on the flank of the relevant animal (see Figure 2.2). Animal-relevant noises were again used to further emphasise the context (six for each species, randomly selected on each trial, each sound 1000–2000 ms in duration).

Procedure

The instructions participants received for this experiment were the same as those of Experiment 1; participants were asked to take on the role of a fictitious veterinarian diagnosing cats or dogs. Participants completed this experiment under either a cat or dog context, but this context was not manipulated. In each of the phases listed below the trial order was randomised between participants.

The training phase was identical to Experiment 1, with each trial beginning with the presentation of the cat or dog image and sound. After 400 ms, the symptoms appeared as a vertical list on the flank of the animal. The order in which symptoms appeared on the vertical list was counterbalanced across trials. Participants diagnosed the animal by pressing one of four keys: “f” for Fips Syndrome, “g” for Gertz Syndrome, “h” for Haust Syndrome and “j” for Jominy Syndrome. After making a response, participants were given corrective feedback for 2000 ms, telling them whether they were correct or incorrect and also what the correct response key and disease were. This feedback was presented on the animal’s flank, replacing the symptoms. If the participant did not respond within 30 s, a time-out message appeared for 2000 ms, informing them that they were out of time and should speed up. Overall, participants received four training phase blocks. The first two blocks consisted of 16 trials, whilst the third and fourth consisted of 32 trials.

After completing the first training phase, participants received a test phase block. Further instructions were given, explaining that they would now be presented with combinations of symptoms that they had not seen before, without corrective feedback. The trial structure of the test phase was identical to the training phase except that, after making a response, instead of receiving corrective feedback, participants now received the message “Response recorded” for 2000 ms. Overall participants received four test phase blocks, each consisting of 96 trials. After completing a test phase block, participants were informed that they would now again be receiving feedback for their responses and received further training. Participants completed four interleaved training-test phase blocks, before finishing the experiment.
3.2. EXPERIMENT 5

Figure 3.1: Participants' accuracy on the abstract training trial types in each training block of Experiment 5. The error bars represent within-subject Cousineau-Morey 95% confidence intervals (Cousineau, 2005; Morey, 2008).

Results

The trial-level raw data, analysis scripts and modelling scripts for this experiment will be available at https://osf.io/hpc6d/ once this experiment is published. A number of the tests below were found to violate Mauchly’s test for sphericity, and so in all cases Greenhouse-Geisser corrections were applied and reported. Figure 3.1 shows the progression of accuracy on the training stimuli over the four training blocks of the experiment. Error bars in this Figure and others within this thesis represent 95% confidence intervals. A 2 x 4 ANOVA was conducted on participants’ accuracy in the training phase using levels of training (16, 32, 64, and 96 trials) and stimulus type (AB vs. AC) as factors. Participants were more accurate in phases where the level of training was higher, $F(3,135) = 71.15, p < .001$, and on common stimuli compared to rare, $F(1,45) = 30.55, p < .001$. There was also a significant interaction between the two factors, $F(3,135) = 6.56, p < .001$, showing that accuracy was greater for AB compared to AC stimuli and that this difference was greater at lower levels of training. Individual t-tests for the difference in accuracy between stimulus types within each training phase revealed significantly greater accuracy for AB over AC stimuli at each training level apart from the final level of 96 training trials.

The proportion of responses to the test stimuli for each of the four test phases can be seen in Table 3.2. The test phases are identified by the number of training trials experienced in total prior to that block. For the key test stimuli, Figure 3.2, shows the progression of rare
responses across test blocks. In the following analyses, data is first analysed within test phase blocks, and then between test phase blocks. All of the within-subject t-test analyses are compared against a level of .5 representing random responding. For each stimulus, a regression-like analysis is conducted. This was performed by fitting a regression line to each individual participant’s response proportions to test stimuli at each level of training (16t, 32t, 64t, 96t). The average gradient of these regression lines was then tested against a null value of 0 to determine if the gradient was significantly different from the null.

No IBRE effect was observed at any training length. After 16 training trials, rare responses to BC were significantly below .5, \( BF = 236.39, t(45) = 4.27, p < .001 \), indicating base-rate following. After further training, rare responding to BC did not differ from .5: 32 trials, \( BF = .20, t(45) = .72, p = .48 \); 64 trials, \( BF = .20, t(45) = .70, p = .49 \); 96 trials, \( BF = .16, t(45) = .21, p = .83 \). The regression-like analysis found that the average gradient for the line of best fit for each participant was significantly greater than 0, \( BF = 10.60, t(45) = 3.12, p = .003 \). In summary, the proportion of rare responding to BC increased as a result of increased

### Table 3.2: Proportion of responses to each of the stimulus types in Experiment 5. Each of the four sub-columns under a response type indicates the number of training trials completed before that specific test phase. Bold type highlights the results of primary theoretical interest. Values within brackets represent response proportions from the simulation of this experiment using the EXIT model with optimised parameters.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>16t</th>
<th>32t</th>
<th>64t</th>
<th>96t</th>
<th>16t</th>
<th>32t</th>
<th>64t</th>
<th>96t</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.52</td>
<td>.57</td>
<td>.59</td>
<td>.61</td>
<td>.58</td>
<td>.61</td>
<td>.21</td>
<td>.19</td>
</tr>
<tr>
<td>AB</td>
<td>.62</td>
<td>.77</td>
<td>.83</td>
<td>.87</td>
<td>.88</td>
<td>.88</td>
<td>.14</td>
<td>.06</td>
</tr>
<tr>
<td>ABC</td>
<td>.56</td>
<td>.55</td>
<td>.51</td>
<td>.24</td>
<td>.46</td>
<td>.38</td>
<td>.33</td>
<td>.46</td>
</tr>
<tr>
<td>AC</td>
<td>.34</td>
<td>.26</td>
<td>.12</td>
<td>.09</td>
<td>.38</td>
<td>.42</td>
<td>.61</td>
<td>.63</td>
</tr>
<tr>
<td>B</td>
<td>.60</td>
<td>.68</td>
<td>.80</td>
<td>.87</td>
<td>.87</td>
<td>.87</td>
<td>.15</td>
<td>.14</td>
</tr>
<tr>
<td>BC</td>
<td>.43</td>
<td>.39</td>
<td>.45</td>
<td>.40</td>
<td>.43</td>
<td>.37</td>
<td>.32</td>
<td>.26</td>
</tr>
<tr>
<td>C</td>
<td>.25</td>
<td>.14</td>
<td>.10</td>
<td>.09</td>
<td>.06</td>
<td>.08</td>
<td>.48</td>
<td>.46</td>
</tr>
<tr>
<td>N</td>
<td>.30</td>
<td>.30</td>
<td>.50</td>
<td>.44</td>
<td>.50</td>
<td>.70</td>
<td>.50</td>
<td>.63</td>
</tr>
<tr>
<td>K</td>
<td>.44</td>
<td>.51</td>
<td>.49</td>
<td>.46</td>
<td>.49</td>
<td>.56</td>
<td>.51</td>
<td>.54</td>
</tr>
</tbody>
</table>

In summary, the proportion of rare responding to BC was significantly greater than 0, \( BF = 10.60, t(45) = 3.12, p = .003 \).
3.2. EXPERIMENT 5

![Figure 3.2](image)

Figure 3.2: Proportion of rare responses to the key test stimuli of Experiment 5. Data points represent the behavioural data, whilst the lines represent the simulated data produced using the optimal parameters for EXIT for this experiment. Test phase blocks are identified according to the total number of training trials completed by that point of the experiment. The error bars represent within-subject Cousineau-Morey 95% confidence intervals (Cousineau, 2005; Morey, 2008).

training, but there was no IBRE in this experiment.

Rare responses to the novel stimulus (N) were significantly greater than .5 in after 16 training trials, $BF = 67.24, t(45) = 3.83, p < .001$ and after 32 training trials, $BF = 32.81, t(45) = 3.56, p = .001$. NHST suggests that rare responses were greater than .5 after 64 training trials, but Bayesian evidence does not support this, $BF = 1.19, t(45) = 2.11, p = .04$. Rare responses to N after 96 training trials did not differ from .5, $BF = .23, t(45) = .89, p = .38$.

The regression-like analysis found that the average gradient for the line of best fit for each participant was not significantly different from 0, but Bayesian evidence does not provide strong evidence for the null, $BF = 0.60, t(45) = 1.69, p = .10$. Overall, the most likely conclusion is that responding to the novel stimulus is preferentially rare at low levels of training, but approximately equally common and rare at higher levels of training.

Common responses to B were significantly greater than rare responses to C after 16 training trials, $BF = 2.42, t(45) = 2.47, p = .02$, although the Bayesian evidence is inconclusive. After 32, 64 and 96 trials of training, common responses to B were equally frequent as rare responses to C: 32 trials, $BF = .19, t(45) = .64, p = .53$; 64 trials, $BF = .22, t(45) = .80, p = .43$; 96 trials, $BF = .16, t(45) = .06, p = .95$.

Common responses to the A stimulus were not significantly different from .5 after 16
training trials, $t(45) = .37, p = .71$, 32 training trials, $t(45) = 1.32, p = .19$, 64 training trials, $t(45) = 1.86, p = .07$, or 96 training trials, $t(45) = 1.62, p = .11$, test phase blocks. The regression-like analysis found that the average gradient for the line of best fit for each participant was not significantly different from 0, $t(45) = 1.28, p = .21$.

It was unclear whether *rare* responding to the blank trials (K) differed from .5 after 16 training trials, $BF = .44, t(45) = 1.48, p = .15$. *Rare* responses did not differ from .5 after 32 training trials, $BF = .17, t(45) = .30, p = .77$, 64 training trials, $BF = .28, t(45) = 1.11, p = .27$, or 96 training trials, $BF = .25, t(45) = .97, p = .34$. The regression-like analysis for K found that the average gradient for the line of best fit for each participant was not significantly different from 0, $BF = .16, t(45) = 0.17, p = .87$. Comparing *rare* responses to K with *rare* responses to N at each training level indicates that the presence or absence of a difference is unclear after 16 training trials, $BF = 1.34, t(45) = 2.17, p = .04$. After 32 training trials, *rare* responses to N were significantly greater than *rare* responses to K, $BF = 20.89, t(45) = 3.39, p = .002$. *Rare* responses to N and K did not differ after 64 training trials, $BF = .33, t(45) = 1.25, p = .22$, or 96 training trials, $BF = .17, t(45) = .25, p = .80$.

**Modelling**

Table 3.3 shows the optimal parameter values, as well as the Root Mean Squared Error (RMSE) and $R^2$ between the simulated and behavioural data when these parameters are given to the model. The optimal parameter values for $P$ and $\phi$ suggest that the cues within the model do not compete for attention substantially and that the model is decisive in its responses. The specificity parameter $c$ is relatively low, suggesting that for the best fit to the behavioural data the model views all the exemplars as similar. The $\sigma$ parameter is also relatively low, indicating that the bias cue does not have much of an effect on responding. The learning rate parameters $\lambda_x$ and $\lambda_w$ are low, whilst the $\lambda_w$ parameter is moderate, suggesting that overall the model learns at a slow rate.

The key response patterns that EXIT should capture are those for the BC, B, C, N and K

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$c$</th>
<th>$P$</th>
<th>$\phi$</th>
<th>$\lambda_x$</th>
<th>$\lambda_w$</th>
<th>$\sigma$</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.701</td>
<td>1.416</td>
<td>3.054</td>
<td>1.000</td>
<td>0.231</td>
<td>1.000</td>
<td>0.227</td>
<td>0.059</td>
</tr>
</tbody>
</table>

*Table 3.3: Optimal parameters for the EXIT model for Experiment 5. RMSE and $R^2$ are calculated between simulated data produced by the model supplied with these parameters and the behavioural data.*

68
cues. In this experiment, responses to BC initially are preferentially common and then shift to .5 at higher training levels. Common responses to B are initially significantly greater than rare responses to C, but change to being equal at higher training levels. Responses to N are initially preferentially rare and then shift to random responding at higher training levels. Finally, responses to K reflect random responding. Values within the brackets in Table 3.2 and the lines in Figure 3.2 represent the simulated data produced by EXIT using the optimal parameters. EXIT manages to capture the response patterns for the BC, B, C and K cues. However, the model does not capture the response pattern for the novel cue until higher levels of training. It is also worth noting that EXIT seems to struggle with BC being close to .5, at least after 32 and 64 training trials. Overall, the RMSE value is very low, indicating a good fit to the data and the $R^2$ is very high, indicating that a large proportion of the variance is captured by the model.

**Discussion**

Experiment 5 did not successfully find the IBRE at any training level. At the lowest level of training, responding to the BC cue reflects base-rate following, whilst at all other training levels responding was at the .5 level. Responding to the novel cue started preferentially rare, as it did in Juslin et al. (2001), and shifted to random responding at higher levels of training. The B>C effect was noted at low levels of training, shifting such that the proportion of common responding to B was equal to the proportion of rare responding to C at higher training levels. Finally, responding to the blank cue was random throughout. The fact that responding to the blank cue does not follow the base rate of the outcomes suggests that participants are not making use of the base rate on a trial where they receive no information.

Much like the experiments in Chapter 2, EXIT was able to successfully capture the response patterns for the IBRE and for the blank cue, K. Further, it also captures the B>C effect at low training levels, something that it struggled to do in the previous chapter. This demonstrates that even without implementing the specific fitting method used in Kruschke (2003), EXIT can capture the B>C effect under certain conditions. However, EXIT was unable to successfully capture the response pattern for the novel cue at low levels of training in this experiment, producing random responding instead of preferentially rare responding.

The fact that Experiment 5 was unable to produce a significant IBRE is problematic. The
response pattern for the BC cue did start to shift towards the classic IBRE of preferentially rare responding as training progressed. It is possible that the reason the IBRE never arose is due to the within-subjects nature of the procedure. Participants experienced 592 trials over the course of this experiment; much more than the experiments in the previous chapter. This could have caused participants to become fatigued, potentially affecting their responding and causing the absence of the IBRE. Experiment 6 seeks to address this issue.

Experiment 6

The experimental procedure of Experiment 6 is very close to that of Experiment 5. The major change is that Experiments 6’s procedure is between-subjects, in order to address the potential confound of fatigue effects noted in Experiment 5. Due to the particularly striking response pattern for the novel cue at the lowest training level of Experiment 5, the training levels in Experiment 6 were adjusted such that the lowest level of training was even lower; to see if the response pattern for the novel cue still arises.

The key goals remain the same as for Experiment 5 in relation to the reduction in the magnitude of the IBRE as a result of training, whilst in terms of the novel cue, this experiment aims to again to produce the result as a test of its reproducibility. EXIT is again fit to the data, although if the response pattern for the novel cue replicates, then it should be unable to capture it.

Method

Participants

95 people participated in this experiment.

Stimuli

The stimuli were presented in the same way as Experiment 5. The same four fictional diseases were used. The combinations of abstract cues and diseases was the same as Table 3.1; with the caveat of there being a single novel cue, $N_1$. The symptom paired with each cue was randomised between participants. Diseases were mapped to the abstract disease types in every possible way between participants.
3.3. EXPERIMENT 6

Procedure

In each of the phases listed below the trial order was randomised between participants. Participants received the same instructions as in Experiment 5. After reading the instructions, participants received a training phase. The trial design of the training phase was the same as that of Experiment 5. The length of the training phases was manipulated between-subjects, resulting in participants being randomly allocated to one of four conditions: 8 trials, 16 trials, 32 trials, or 64 trials. As 8 trials is not enough to fully counterbalance the position of symptoms (top/bottom), the position was randomised for each trial for the 8-trial condition.

After completing the training phase, participants received instructions for the test phase as in Experiment 5. The test phase was largely the same as in Experiment 5, although participants received only a single test phase.

Results

The trial-level raw data, analysis scripts and modelling scripts for this experiment will be available at https://osf.io/xhdeu/ once this experiment is published. One participant’s data was discarded due to them failing to respond to the blank cue trials. Participants received one of four conditions: 8 trials, 16 trials, 32 trials or 64 trials. The number of participants in each condition was 24, 23, 23 and 24 respectively. Figure 3.3

Figure 3.3: Participants’ accuracy on the abstract training trial types in each condition of Experiment 6. Accuracy represents the percentage of correct answers in the final 8 training trials. The error bars represent difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).
3.3. EXPERIMENT 6

shows the accuracy of participants on the training stimuli within each condition.

Looking at participants’ accuracy in the training phase, accuracy was higher in conditions where they received greater amounts of training, $F(3, 90) = 11.78, p < .001$, and greater for the common stimuli than the rare stimuli, $F(1, 90) = 22.08, p < .001$. No significant interaction between these factors was found. Individual t-tests found greater accuracy on the common stimuli than the rare stimuli in the 16t and 32t conditions but not the 8t or 64t conditions.

The proportion of responses to the test phase stimuli in each of the four conditions can be seen in Table 3.4. For the key test stimuli, Figure 3.4, shows the proportion of rare responses in each condition’s test phase. In the following analyses, data is first analysed within condition, and then between conditions. All of the within-subject t-test analyses are compared against a level of .5 representing random responding. For each stimulus a linear regression analysis is conducted, using response proportion as the dependent variable and level of training (8, 16, 32, or 64 trials) as the independent variable.

Table 3.4: Proportion of responses to each of the stimulus types in Experiment 6. Each of the four sub-columns under a response type indicates the specific condition. Bold type highlights the results of primary theoretical interest. Values within brackets represent response proportions from the simulation of this experiment using the EXIT model with optimised parameters.
3.3. EXPERIMENT 6

Looking first at the BC stimulus, rare responses were significantly lower than .5 after 8 trials of training, $BF = 8.05, t(23) = 3.07, p = .01$, and after 16 trials of training, $BF = 5.88, t(22) = 2.91, p = .01$. The proportion of rare responses did not differ from .5 after 32 trials of training, $BF = .25, t(22) = .52, p = .61$, or 64 trials of training, $BF = .26, t(23) = .68, p = .50$. Simple linear regression, aiming to predict response proportion based on training length, found a significant regression equation, $BF = 9.30, F(1,92) = 8.74, p = .004$, with an adjusted $R^2$ of 0.08. Condition significantly predicted response proportion, $a = .32, b = .004, t(92) = 2.96, p = .004$. As in Experiment 5, there was no evidence for an IBRE although, as before, the proportion of rare responses to BC increased with greater amounts of training.

Rare responses to the N stimulus were significantly greater than .5 in the 8-trial condition, $BF = 5.50, t(23) = 2.87, p = .01$. NHST suggests this is also the case in the 16-trial condition, $BF = 2.74, t(22) = 2.50, p = .02$, condition, although the Bayesian evidence is inconclusive. Rare responses did not differ from .5 in the 32-trial condition, $BF = .30, t(22) = .85, p = .41$, nor the 64-trial condition, $BF = .22, t(23) = .18, p = .86$. Simple linear regression found a significant regression equation, $BF = 2.58, F(1,92) = 5.70, p = .02$, with an adjusted $R^2$ of 0.05, although the Bayesian evidence was inconclusive. Condition was a significant predictor of response proportion, $a = .71, b = -.04, t(92) = 2.39, p = .02$.

Common responses to B were significantly greater than rare responses to C in the 16-trial condition, $BF = 4.18, t(22) = 2.03, p = .05$. NHST suggests this is also the case in the 32-trial condition, $BF = .26, t(22) = .50, p = .62$, although the Bayesian evidence is inconclusive.
condition, $BF = 17.88, t(22) = 3.47, p = .002$. The Bayesian evidence for the presence of a difference was inconclusive in the 8-trial condition, $BF = .59, t(23) = 1.52, p = .14$, inconclusive in the 32-trial condition, $BF = .52, t(22) = 1.40, p = .17$, and inconclusive in the 64-trial condition, $BF = .46, t(23) = 1.31, p = .20$.

**Common** responses to the A stimulus were not significantly different from .5 in the 8-trial condition, $t(23) = 1.27, p = .22$, nor the 16-trial condition, $t(22) = 1.06, p = .30$, nor the 32-trial condition, $t(22) = 1.98, p = .06$. **Common** responses were significantly greater than .5 in the 64-trial condition, $t(23) = 5.90, p < .001$, condition. Simple linear regression found a significant regression equation, $F(1,92) = 14.92, p < .001$, with an adjusted $R^2$ of 0.13. Condition significantly predicted response proportion, $a = .44, b = .01, t(92) = 3.86, p < .001$.

**Rare** responses to the blank K stimulus did not differ from .5 in the 8-trial condition, $BF = .27, t(23) = .68, p = .50$, nor the 16-trial condition, $BF = .29, t(22) = .81, p = .43$. The Bayes evidence for the presence or absence of a difference was inconclusive in the 32-trial condition, $BF = .34, t(22) = .98, p = .34$, and inconclusive in the 64-trial condition, $BF = .46, t(23) = 1.31, p = .20$. Simple linear regression did not find a significant regression equation, $BF = .42, F(1,92) = 1.50, p = .22$, with an adjusted $R^2$ of 0.01; although the Bayesian evidence was inconclusive. Condition was not a significant predictor of response proportion, $a = .51, b = -.002, t(92) = 1.23, p = .22$.

Comparing **rare** responses to K to **rare** responses to N at each training level indicated that response rates did not differ in the 32-trial condition, $BF = .22, t(22) = .09, p = .93$, nor in the 64-trial condition, $BF = .26, t(23) = .62, p = .54$. The Bayesian evidence for the presence or absence of a difference was inconclusive for the 8-trial condition, $BF = 1.15, t(23) = 1.99, p = .06$. In the 16-trial condition, **rare** responses to N were significantly greater than **rare** responses to K, $BF = 13.65, t(22) = 3.34, p = .003$.

**Modelling**

Table 3.5 shows the optimal parameter values, as well as the Root Mean Squared Error

<table>
<thead>
<tr>
<th>Parameters</th>
<th>c</th>
<th>P</th>
<th>$\phi$</th>
<th>$\lambda_x$</th>
<th>$\lambda_w$</th>
<th>$\lambda_y$</th>
<th>$\sigma$</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.588</td>
<td>2.332</td>
<td>4.006</td>
<td>1.000</td>
<td>0.127</td>
<td>1.000</td>
<td>0.052</td>
<td>0.062</td>
<td>0.933</td>
</tr>
</tbody>
</table>
(RMSE) and $R^2$ between the simulated and behavioural data when these parameters are given to the model. The optimal parameters are similar to those of Experiment 5’s optimal parameter set. The $P$ and $\phi$ parameters indicate little competition for attention between cues and that the novel is decisive in its responses respectively. The $c$ parameter is again low reflecting increased confusability between exemplars within the model. The $\sigma$ parameter is low, indicating that the bias cue within the model again has very little influence. $\lambda_w$, $\lambda_v$, and $\lambda_x$ are low, reflecting generally low learning rates for the model.

In this experiment, responses to BC initially are preferentially common and then shift to random responding at higher training levels. Common responses to B are initially significantly greater than rare responses to C, but change to being equal at higher training levels. Responses to N are initially preferentially rare and then shift to random responding. Finally, responses to K reflect random responding throughout. Values within the brackets in Table 3.4 and the lines in Figure 3.4 represent the simulated data produced by EXIT using the optimal parameters. EXIT again manages to capture the response patterns for the BC, B, C and K cues. However, the model continues to not capture the response pattern for the novel cue until higher levels of training. Overall, the RMSE value is very low, indicating a good fit to the data and the $R^2$ is very high, indicating that a large proportion of the variance is captured by the model.

**Discussion**

Experiment 6 again did not observe a significant IBRE at any level of training, with the pattern of results again initially reflecting base rate following before shifting to random responding. Responding to the novel cue was again rare at low levels of training, before shifting to random responding at higher training levels. The $B>C$ effect was again observed at the second lowest training level, shifting to equal proportions of common responding to B and rare responding to C at higher levels of training. Responding to the blank cue was again random.

Much like Experiment 5, EXIT is able to capture the response patterns for the BC, blank, B and C cues. Unsurprisingly, it is still unable to capture the pattern of responding for the novel cue. The pattern of rare responding to the novel cue at low levels of training is replicated in this experiment. However, the issue of not finding an IBRE is odd. As the IBRE reported in Juslin et al. (2001) is not supported by statistical tests, it is possible that they do not find the IBRE in their study. This would explain the contradictory results of their
study and the experiments within this chapter. To address this, Experiment 7 implements both the very low level of training used in this experiment and the very large level of training used in Experiments 3 and 4 of Chapter 2.

**Experiment 7**

Whilst the IBRE was not found in Experiments 5 or 6, it is possible that participants were simply not trained enough to produce the effect. Experiment 7 addresses this by keeping the between-subject procedure of Experiment 6, but simplifying it to two training level conditions. The first is the lowest training level used in Experiment 6, whilst the second is based on the largest number of trials it took participants in Experiment 1 to pass criterion. Even if the IBRE is not observed in the low level of training condition, the high level of training condition should produce the effect; given the results of Experiments 3 and 4. EXIT is again fit to the behavioural data for this experiment. The key outcomes for this experiment are to find an IBRE at least in the high level of training condition, and to again test EXIT’s ability to produce the behavioural response patterns.

**Method**

**Participants**

43 people participated in this experiment.

**Stimuli**

The stimuli were presented in the same way as Experiments 5 and 6. The same four fictional diseases were used. The abstract stimuli types used in this experiment were the same as Experiment 5 (see Table 3.1); with the caveats that there was a single novel cue, \( N_1 \), and that, due to a technical error, no blank cue, \( K_1 \), was used. The symptom paired with each cue was randomised between participants. Diseases were mapped to the abstract disease types in every possible way between participants. Animal-relevant noises were not used in this experiment.

**Procedure**

The experimental procedure of Experiment 7 was the same as that of Experiment 6; with the only difference being the length of training participants experienced in each condition. Participants received either 8 trials of training or 192 trials. Participants again received a single training phase followed by a single test phase.
3.4. EXPERIMENT 7

Figure 3.5: Participants’ accuracy on the abstract training trial types in each condition of Experiment 7. Accuracy represents the percentage of correct answers in the final 8 training trials. The error bars represent difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

Results

The trial-level raw data, analysis scripts and modelling scripts for this experiment will be available at https://osf.io/xgdj5/ once this experiment is published. Participants were randomly assigned to one of the two conditions, 8 training trials or 192 training trials. This resulted in 21 and 22 participants in each condition respectively. Figure 3.5 shows the accuracy of participants on the training stimuli within each condition.

Looking at participants’ accuracy on the final 8 trials in the training phase, accuracy was significantly greater in the 192-trial condition than the 8-trial condition, $F(1, 41) = 143.53, p < 0.001$ and on AB than AC stimuli, $F(1, 41) = 8.84, p = .001$. There was evidence of significant interaction between these two factors, $F(1, 41) = 8.38, p = 0.001$. Individual t-tests for the difference in accuracy between stimulus types within each training phase revealed significantly greater accuracy for AB over AC stimuli in the 8t but not the 192t condition.

The proportion of responses to the test phase stimuli can be seen in Table 3.6. In the following analyses, data is first analysed within condition, and then between condition. All of the within-subject t-test analyses are compared against a level of .5 representing random responding.
Table 3.6: Proportion of responses to each of the stimulus types in Experiment 7. The two sub-columns indicate the specific condition. Bold type highlights the results of primary theoretical interest. Values within brackets represent response proportions from the simulation of this experiment using the EXIT model with optimised parameters.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>Common 8t</th>
<th>Rare 8t</th>
<th>Common Other 8t</th>
<th>Rare Other 8t</th>
<th>Common 192t</th>
<th>Rare 192t</th>
<th>Common Other 192t</th>
<th>Rare Other 192t</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.48 (.42)</td>
<td>.70 (.70)</td>
<td>.18 (.23)</td>
<td>.26 (.19)</td>
<td>.17 (.18)</td>
<td>.03 (.05)</td>
<td>.17 (.17)</td>
<td>.01 (.06)</td>
</tr>
<tr>
<td>AB</td>
<td>.56 (.58)</td>
<td>.92 (.96)</td>
<td>.16 (.16)</td>
<td>.07 (.01)</td>
<td>.15 (.13)</td>
<td>.00 (.01)</td>
<td>.13 (.13)</td>
<td>.01 (.02)</td>
</tr>
<tr>
<td>ABC</td>
<td>.45 (.50)</td>
<td>.35 (.35)</td>
<td>.20 (.22)</td>
<td>.63 (.62)</td>
<td>.17 (.14)</td>
<td>.00 (.01)</td>
<td>.18 (.14)</td>
<td>.02 (.02)</td>
</tr>
<tr>
<td>AC</td>
<td>.36 (.32)</td>
<td>.07 (.02)</td>
<td>.25 (.31)</td>
<td>.93 (.95)</td>
<td>.20 (.19)</td>
<td>.00 (.02)</td>
<td>.19 (.18)</td>
<td>.00 (.01)</td>
</tr>
<tr>
<td>B</td>
<td>.49 (.46)</td>
<td>.93 (.89)</td>
<td>.21 (.18)</td>
<td>.04 (.03)</td>
<td>.15 (.18)</td>
<td>.00 (.03)</td>
<td>.15 (.18)</td>
<td>.03 (.05)</td>
</tr>
<tr>
<td>BC</td>
<td>.37 (.39)</td>
<td>.31 (.26)</td>
<td>.29 (.24)</td>
<td>.69 (.66)</td>
<td>.14 (.19)</td>
<td>.00 (.04)</td>
<td>.20 (.18)</td>
<td>.00 (.04)</td>
</tr>
<tr>
<td>C</td>
<td>.18 (.22)</td>
<td>.02 (.01)</td>
<td>.41 (.33)</td>
<td>.97 (.94)</td>
<td>.19 (.23)</td>
<td>.00 (.03)</td>
<td>.22 (.22)</td>
<td>.01 (.02)</td>
</tr>
<tr>
<td>N</td>
<td>.39 (.50)</td>
<td>.52 (.51)</td>
<td>.61 (.50)</td>
<td>.48 (.49)</td>
<td>.00 (.00)</td>
<td>.00 (.00)</td>
<td>.00 (.00)</td>
<td>.00 (.00)</td>
</tr>
</tbody>
</table>

Looking first at the 8-trial condition, rare responses to the BC stimulus were significantly below .5, \( BF = 32.08, t(20) = 3.79, p = .001 \), representing base rate responding. For the novel stimulus, N, the Bayesian evidence for the presence or absence of a difference from .5 was inconclusive, \( BF = .49, t(20) = 1.32, p = .20 \). Similarly, the Bayesian evidence for the presence or absence of a difference between the common response proportion to the B stimulus and the rare response proportion to the C stimulus was also inconclusive, \( BF = .35, t(20) = 0.99, p = .34 \). The proportion of common responses to the A stimulus was not significantly different from .5, \( t(20) = 0.34, p = .74 \).

![Figure 3.6](image-url)  
*Figure 3.6: Proportion of rare responses to the key test stimuli of Experiment 7. Data points represent the behavioural data, whilst the lines represent the simulated data produced using the optimal parameters for EXIT for this experiment. The error bars represent difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).*
3.4. EXPERIMENT 7

Looking secondly at the 192-trial condition, rare responses to the BC stimulus were significantly higher than .5, $BF = 7.75, t(21) = 3.06, p = .006$, indicating the presence of an IBRE. Rare responses to the novel N stimulus did not differ from .5, $BF = .23, t(21) = 0.17, p = .87$. The Bayesian evidence for the $B \rightarrow \text{common}$ versus $C \rightarrow \text{rare}$ difference was inconclusive, $BF = .56, t(21) = 1.45, p = .16$. The proportion of common responses to the A stimulus was significantly greater than .5, $t(21) = 3.82, p < .001$.

There were significantly more rare responses to the BC stimulus in the 192-trial condition than the 8-trial condition, $BF = 803.96, t(41) = 4.80, p < .001$. The corresponding comparison for the novel stimuli (N) was inconclusive, $BF = .45, t(41) = 1.02, p = .31$. There was greater common responding to the B stimulus in the 192-trial condition than the 8-trial condition, $BF = 97095.63, t(28) = 6.33, p < .001$ and greater rare responding to the C stimulus, $BF = 47928204, t(21) = 8.32, p < .001$. There was significantly greater common responding to the A stimulus in the 192-trial condition than the 8-trial condition, $t(37) = 2.55, p = .02$.

**Modelling**

Table 3.7 shows the optimal parameter values, as well as the Root Mean Squared Error (RMSE) and $R^2$ between the simulated and behavioural data when these parameters are given to the model. The optimal parameters for Experiment 7 are substantially different from those of Experiments 5 and 6. The c parameter has increased suggesting that the exemplars for the model are less confusable with each other for this simulation. The P and φ parameters have also increased relative to the previous two experiments, suggesting further decreased cue competition for attention and increased decisiveness in the responses of the model. $\lambda_g$ and $\lambda_x$ parameters have increased reflecting greater learning rates within the model. However $\lambda_w$ has remained low, suggesting that the learning of associative weights within the model is not a priority. The $\sigma$ parameter is close to its lower limit, again limiting the influence of the bias cue.

In this experiment, responses to BC initially are preferentially common and then shift to

<table>
<thead>
<tr>
<th>Parameters</th>
<th>c</th>
<th>P</th>
<th>φ</th>
<th>$\lambda_g$</th>
<th>$\lambda_w$</th>
<th>$\lambda_x$</th>
<th>$\sigma$</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.772</td>
<td>3.801</td>
<td>4.498</td>
<td>2.910</td>
<td>0.091</td>
<td>1.795</td>
<td>0.031</td>
<td>0.038</td>
<td>0.979</td>
</tr>
</tbody>
</table>
preferentially rare at the higher training level. Common responses to B are indistinguishable from rare responses to C. Finally, responses to N are initially preferentially rare and then shift to random responding. Values within the brackets in Table 3.6 and the lines in Figure 3.6 represent the simulated data produced by EXIT using the optimal parameters. EXIT again manages to capture the response patterns for the BC cue. However, the model continues to not fully capture the response pattern for the novel cue. Overall, the RMSE value is very low, indicating a good fit to the data and the $R^2$ is very high indicating that a large proportion of the variance is captured by the model.

Discussion

Experiment 7 succeeded in finding a significant IBRE; but only in the high level of training condition. Responding to the novel N cue was preferentially rare in the low level of training condition, shifting to random responding in the high level of training condition. Perhaps unsurprisingly for this experiment, common responding to B was not greater than rare responding to C at any level of training. This is most likely due to participants not having learnt enough to show a difference in the low training level condition and having learn too much to show a difference in the high training level condition.

As it has in the previous two experiments, EXIT captures the response pattern for the BC, B and C cues, but still cannot capture the preferentially rare response pattern for the novel cue in the low training level condition.

Chapter Discussion

The results of the experiments within this chapter reveal a number of important points. Contrary to the results of Juslin et al. (2001) it seems that producing a significant IBRE requires a certain level of training. The pattern of responding for the novel cue reported in Juslin et al. (2001) does seem accurate, with preferentially rare responding at low levels of training. The blank cue used in Experiments 5 and 6 seems to be responded to at random; rather than showing common responding reflecting the base rate of the outcomes. The B>C effect seems to occur at low levels of training, further agreeing with the results of Juslin et al. (2001).

Whilst these experiments do not support the IBRE findings noted in Juslin et al. (2001), the pattern for the novel cue noted in that study and the work of Johansen et al. (2007) does seem accurate. This is particularly relevant given that the EXIT model cannot capture
this result. As EXIT is the best current formal model able to capture the IBRE, a primary part of this thesis is the assessment of the models’ capability to capture behavioural response patterns in IBRE experiments. The fact that it cannot capture preferentially rare responding to a novel cue is a major issue. As mentioned in the discussion of Experiment 5, it seems that there is no mechanism within the architecture of EXIT that could produce this response pattern.
Chapter 4

Neural Correlates of the Inverse Base Rate Effect

Introduction

Technological developments within the last half-century have led to an increased number of neuroimaging methodologies used in the investigation of human learning. Both Electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) have been used to investigate learning phenomena (e.g. Eippert, Gamer, & Büchel, 2012; Tobler, O’Doherty, Dolan, & Schultz, 2006; Wills et al., 2014). However, previous work making use of neuroscience methodologies in IBRE procedures is limited; to my knowledge only two published papers directly examine the IBRE with such methodologies.

Wills et al. (2014) report the first IBRE study using an electrophysiological methodology, making use of Event Related Potentials (ERP). They identify previously reported potentials related to selective attention (Anllo-Vento & Hillyard, 1996), including Selection Negativity (SN), expecting to find evidence of these potentials in their study. Specifically, they predict an SN for the C cue relative to the B cue. They further predict that this difference would not occur for cues that are presented at the same frequency as B and C, but are not presented in compound with a shared cue such as A. These predictions were found to be correct, supporting an error-driven attentional account of the IBRE. They found both a behaviourally significant IBRE and a significant B>C effect, discounting the possibility that these potentials were driven by a difference in learning between B and C. These results determine the time frame of the brain activations that correlate with the IBRE, but due to the methodological limitations of ERP, specific cortical and sub-cortical regions within the brain associated with the IBRE cannot be identified. As EEG measures the change in electrical activity on the scalp as a result of brain activations, it cannot be used to determine the specific spatial location of neural generators within the brain itself. There
are certain analysis techniques that attempt to solve this issue of source localisation including low-resolution brain electromagnetic tomography (LORETA; Pascual-Marqui, Michel, and Lehmann (1994)) and dipole fitting (De Munck, van Dijk, & Spekreijse, 1988)). However, these techniques do have limitations in terms of their ability to capture activations in deep brain regions. LORETA does not include deep brain areas such as the striatum in its solution space, while dipole fitting suffers from the inverse problem; where the is the possibility of an infinite number of possible solutions. In summary, whilst EEG can determine the origin of activations in superficial brain areas, a more spatially accurate approach for determining the origin of activations throughout the brain is fMRI.

The second neuroscience study of the IBRE, O'Bryan et al. (2018), makes use of an fMRI methodology. fMRI is capable of capturing cortical and sub-cortical activations in specific brain areas, providing the ability to capture the origin of brain activations in relation to phenomena such as the IBRE. The technique is based on measuring Blood Oxygen Level Dependent (BOLD) signals. BOLD represents the idea that blood has differing magnetic properties based on whether or not it is oxygenated. Oxygenated blood is diamagnetic, while deoxygenated blood is paramagnetic. These differing magnetic properties are detectable using MRI. When neuronal activity increases in a brain region, oxygen use increases to provide energy. This causes a decrease in blood oxygenation, and so an increase in deoxygenated blood. As a result of this, the paramagnetic signal from that brain region increases. It is this change in signal that is then measured. However, as this change is not instantaneous, fMRI struggles to identify the time frame of these activations. fMRI is therefore a complementary methodology to ERP.

The key behavioural result for the IBRE in O'Bryan et al. (2018) was that participants were significantly more likely to respond rare to BC when tested against a chance level of .25, a value which reflects random responding. However, their analysis method is inherently flawed. They make use of four outcomes in their procedure; hence the chance level of .25. This seems logical; however two outcomes are rare and two are common. This is due to them doubling up the classic design of AB leading to common and AC leading to rare such that each of the cues have two instantiations e.g. A1 and A2. Looking at the response proportions for the A1, B1 and C1 cues when presented alone in O'Bryan et al. (2018), it is clear that their participants know the specific common and rare outcomes associated with those cues. This would mean that there are only two outcomes relevant on those trials; one common and one rare. Therefore, the chance level reflecting random responding
would be .5 instead of .25. The rare response proportion to the BC cue in O’Bryan et al. (2018) is close to .5, and so it is unlikely that when tested against a .5 chance level that this proportion would be significantly greater. This would suggest that contrary to what they report, O’Bryan et al. (2018) do not find a significant IBRE. It would be premature to base future work on these results, and so a further investigation of the IBRE within an fMRI methodology is merited.

The aims of this chapter are threefold. First; to produce a significant IBRE within an fMRI methodology. Second; to identify activations that correlate with the IBRE within a number of pre-defined brain regions linked to prediction error; a major component of the error-driven learning account implemented within Kruschke (2001a)’s EXIT model. These regions include the striatum, the dorsolateral prefrontal cortex, the medial prefrontal cortex and the anterior cingulate cortex. Third; to evaluate the changes in activation in these regions between key individual stimulus test trials. Previous work (Wills et al., 2014) suggests that there should be greater activations for the C cue compared to the B cue, and that there should be no difference in activation for frequency matched cues (D and E), trained in the absence of a shared cue.
Experiment 8

Experiment 3 of Chapter 2 was a direct replication of the behavioural procedure of Wills et al. (2014). Whilst the evidence for an IBRE in Experiment 3 was inconclusive, an IBRE was found in the same-context condition of Experiment 4; which used the same procedure. As such, this procedure is able to find the IBRE, as well as the $B>C$ effect. Experiment 8 uses this procedure within an fMRI methodology. Whilst the behavioural analysis follows that of Experiment 4, both whole-brain and Region of Interest (ROI) analysis is used to identify activations in specific brain regions previously identified as related to attentional learning tasks. The predictions for Experiment 8 follow those of Wills et al. (2014), predicting a difference in brain activations for the C and B test cues and no difference in brain activations for the frequency matched cues that do not share a cue at training. Further, analysis of the difference of these differences should show a greater difference in activations for the C and B test cues relative to the difference in activations of the frequency matched cues. This is the critical analysis; whilst previous work suggests that the first two comparisons should differ, this provides a direct test of this predicted difference.

Method

Participants

34 people participated in this experiment.

Figure 4.1: An example trial presentation from the training phase of Experiment 8. Image author: Anna Robertson, CC-BY-SA 4.0
Stimuli

The stimuli were the same as those used in Experiment 3. The stimuli were displayed on a back-projection screen positioned at the foot end of the MRI scanner and viewed via a mirror mounted on a head coil. The size of the “cells” was increased to 250 x 250 pixels. This was done to ensure participants could see the stimuli clearly on the projector screen.

An example of the cells on an experimental trial can be seen in Figure 4.1.

The stimuli were again assigned at random to one of 7 abstract cues (A-G) for each participant. As before, each abstract cue had 3 stimuli assigned to it, resulting in 21 different “cells” being used. Table 4.1 shows the combinations of abstract cues and diseases presented in the training and test phases of the experiment. Two diseases were used as outcomes in this experiment: “Jominy Fever” and “Phipps Syndrome”. Diseases were mapped to the abstract disease types in both possible ways, across participants.

Procedure

In each phase of this experiment the trial order was randomised. Participants were given instructions before entering the MRI scanner (1.5 T imager). Participants were asked to take on the role of a doctor, diagnosing patients with either “Jominy Fever” or “Phipps Syndrome” on the basis of the “cells” they were presented with. The response key that represented each disease was explained to the participant before the task, and was fully counterbalanced between participants. The disease that was abstractly common or rare was also fully counterbalanced between participants.

Button-press responses and RTs were measured using a fibre-optic button box (4-Button...
Figure 4.2: The structure of a trial in Experiment 8. The variable duration fixation cross was displayed from 0 ms. Its minimum duration was 250 ms and its maximum duration was 3500 ms. The rest of the trial proceeded immediately after the presentation of the fixation cross. The view box was presented alone for 500 ms, then remained onscreen for a further 2000 ms. The “cell(s)” is/are displayed for 2000 ms after the view box was presented alone. The feedback was displayed for 500 ms after this.

Inline Fiber Optic Response Pad). The training phase consisted of 10 blocks of 36 trials; 360 trials in total. Figure 4.2 shows the structure of a trial in Experiment 8. Trials varied in length depending on a variable duration fixation cross presented in the centre of the screen. The durations were generated using an exponential distribution, using the method described in Haberg, Zito, Patria, and Sanes (2001). The range of the duration was 250 ms - 3500 ms with a mean duration of 1284 ms. After this, a grey view-box (see Figure 4.1) was displayed on its own for 500 ms to give participants an idea of where the cells would appear. The “cells” appeared to the top and bottom of the view-box and the “cell” that was presented towards the top and bottom was randomised on each trial. The “cells” remained on screen for a maximum duration of 2000 ms, during which time participants made their diagnosis using either the left or right button on the button box. After this, participants received corrective feedback telling them if they were right or wrong and what the correct diagnosis was for 500 ms. If a response was not made during the 2000 ms, then participants instead received a time-out message for 500 ms telling them they were out of time and to speed up.

Further instructions were given at the start of the test phase. Participants were informed that they would continue to diagnose patients and would see some “cells” that they had seen before. They were told that they would also see novel “cells” and “cell” combinations that they would not receive feedback for. The test phase consisted of 282 trials in total. The trial structure in the test phase was the same as in the training phase, but with the addition of single “cells” being presented in the centre of the view-box. The variable duration fixation point at the start of the trial maintained its range but had an average duration of
Participants continued to receive corrective feedback for the trial types they saw in the training phase to maintain learning. On trials for which participants did not receive feedback, they instead received the message “DATA MISSING” and a series of question marks. After completing the test phase, participants finished the experiment.

**fMRI Data Acquisition**

Images were collected using a 1.5-T Gyroscan magnet equipped with a Sense coil (Philips, Amsterdam, The Netherlands). A T2*-weighted echo-planar sequence was used (repetition time = 3000 ms, echo time = 45 ms, flip angle = 90°, 32 transverse slices, field of view = 240 mm, 3.5 × 2.5 × 2.5 mm). The training phase comprised two runs of 242 scans, and the test phase two runs of 187 scans. Standard volumetric anatomical MRI was performed after functional scanning by using a 3-D T1-weighted pulse sequence (repetition time = 25 ms, echo time = 4.1 ms, flip angle = 30°, 160 axial slices, 1.6 × 0.9 × 0.9 mm).

**Analysis of fMRI Data**

Analyses were carried out using SPM12 software (The FIL Methods Group, 2014). Functional images were corrected for acquisition order, realigned to the mean image, and resliced to correct for motion artifacts. The realigned images were coregistered with the structural T1 volume, and the structural volumes were spatially normalized. The spatial transformation was applied to the realigned T2* volumes, which were spatially smoothed using a Gaussian kernel of 8 mm FWHM. Data were high-pass filtered (1/128 Hz) to account for low-frequency drifts. The BOLD response was modelled (GLM) by a canonical hemo-dynamic response function with temporal and dispersion derivatives.

In the individual participant models, the critical trials for comparisons (B, C, D, E) were included as individual regressors, with the other trial types and time outs included as two further separate regressors. The duration of each event was modelled as the participant’s RT for that trial, an approach advocated in Grinbrand, Erdeniz, Lindquist, Ferrera, and Hirsch (2008). Three analyses were completed; comparing C-B, comparing E-D and the critical analysis, comparing the levels of activation in the previous two comparisons. These comparisons were a block design, using individual stimulus scans averaged across all presentations in the test phase, i.e. all presentations of C vs. all presentations of B. Whilst training phase scans were completed, analysis of singular stimuli avoids the issue

‡‡ The slight difference in the average duration is due to the differing number of trials participants received during the training and test phases.
of determining attentional allocation in the compound stimuli used in the training phase. It also avoids the issue of differing participant learning rates that could affect the results if the training data were used for analysis. The whole brain analyses were completed using a general linear model with a combined statistical threshold of $p < .001$ (uncorrected) and a threshold of 106 contiguous voxels, which together produce an overall corrected threshold of $p < .05$. These values were estimated using AlphaSim as implemented in the REST toolbox (Song et al., 2011). For the three imaging analyses conducted, smoothness estimates of [9.8, 9.8, 9.5], [9.8, 9.8, 9.6] and [9.5, 9.5, 9.2] mm were used respectively (these were a group level estimate calculated in SPM12 using the group residuals from the general linear model).

To address the potential for areas in which activations are noted to change across the experiment, a more conservative method of ROI analysis was used, where the regions included in the analysis were pre-determined instead of at the individual participant level, based on regions activated for individual participants. These regions were drawn from studies investigating the behavioural processes of error-driven learning; specifically prediction error. Prediction error is a part of error-driven learning theories (e.g., Kruschke, 2001a) in that it is assumed that attention is reallocated to stimuli in order to reduce the future occurrence of error. In the case of the IBRE, these theories predict greater attention being paid to C than B in the case of the stimulus compound BC in an attempt to reduce prediction error and assume this drives the IBRE. Specific regions that were linked to prediction error and were presumed to be differentially activated were the ventral and dorsal striatum (e.g. Fouragnon, Retzler, & Philiastides, 2018; Garrison, Erdeniz, & Done, 2013; McClure, Berns, & Montague, 2003; O’Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Pagnoni, Zink, Montague, & Berns, 2010; Park et al., 2010), the right dorsolateral prefrontal cortex (BA 9 and BA 46; e.g. Fletcher et al., 2001; Fouragnon et al., 2018; Spoormaker et al., 2011; Turner et al., 2004), medial prefrontal cortex (BA 9 and BA 10; e.g. Fouragnon et al., 2018; Spoormaker et al., 2011) and anterior cingulate cortex (BA 24, BA 32 and BA 33; e.g. Botvinick, Braver, Barch, Carter, & Cohen, 2001; Critchley, Tang, Glaser, Butterworth, & Dolan, 2005; Hoffman & Falkenstein, 2011; Rodriguez, 2009).

The mask used for the ROI analysis was constructed using the WFU Pickatlas (Maldjian, Laurienti, Burdette, & Kraft, 2003), and comprised the bilateral caudate, putamen and nucleus accumbens, as well as BA’s 9, 10, 24, 32, 33 and 46. The number of voxels within this mask was 11952. For these analyses, a general linear model was again
used with a combined statistical threshold of $p = .005$ and 28 contiguous voxels, which
together produce an overall corrected threshold of $p < .05$, as estimated by AlphaSim.
The smoothness estimates were the same as used for the whole-brain analysis. For all
analyses, normalized MNI space coordinates were transformed to Talairach space using
GingerALE (Eickhoff et al., 2011) and assigned anatomical labels using TalairachClient
(http://talairach.org/client.html) as per the atlas of Talairach and Tournoux (1988).

Results

Behavioural Analyses

The trial-level raw data and analysis scripts for this experiment will be available at
https://osf.io/yw6fj/ once this experiment is published. Similarly the raw imaging
data will also be made available in this archive. Five participants’ data were removed due
to excessive head movements during the experiment, rendering their fMRI data unusable
or incomplete. Participants’ accuracy in the final block of training was then assessed
under the same criterion for learning as in Experiment 3; where participants scoring less
than 72% were removed (Wills et al., 2014). Applying this criterion here necessitated
the removal of 7 participants. Three of these participants also displayed excessive head
movements, and so had already been excluded. After applying the above exclusions, the
final data set contained 25 participants.

The accuracy of participants across the training phase can be seen in Figure 4.3. Looking
4.2. EXPERIMENT 8

Table 4.2: Proportion of responses to each of the stimulus types presented in the test phase of Experiment 8. Bold type highlights the results of primary theoretical interest.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.76</td>
<td>.24</td>
</tr>
<tr>
<td>AB</td>
<td>.92</td>
<td>.08</td>
</tr>
<tr>
<td>AC</td>
<td>.19</td>
<td>.81</td>
</tr>
<tr>
<td>B</td>
<td>.92</td>
<td>.08</td>
</tr>
<tr>
<td>BC</td>
<td>.35</td>
<td>.65</td>
</tr>
<tr>
<td>C</td>
<td>.15</td>
<td>.85</td>
</tr>
<tr>
<td>D</td>
<td>.85</td>
<td>.15</td>
</tr>
<tr>
<td>DE</td>
<td>.44</td>
<td>.56</td>
</tr>
<tr>
<td>E</td>
<td>.24</td>
<td>.76</td>
</tr>
<tr>
<td>FD</td>
<td>.96</td>
<td>.04</td>
</tr>
<tr>
<td>GE</td>
<td>.11</td>
<td>.89</td>
</tr>
</tbody>
</table>

first at accuracy in the first and final training blocks, accuracy in the final block was significantly higher than the first block, \( F(1, 24) = 324.63, p < .001 \). No other significant main effects or interactions were found. Looking at the final block of training alone, accuracy was significantly higher for the common stimuli compared to the rare stimuli, \( F(1, 24) = 5.23, p = .03 \). No other significant main effect or interaction between these two factors was found.

Table 4.2 shows the response proportions for each of the stimuli presented in the test phase. The IBRE test stimulus BC was found to have a significantly greater proportion of rare responses than .5, \( BF = 6.27, t(24) = 2.93, p = .01 \), indicating the presence of an IBRE. The Bayes evidence for the presence or absence of a difference between common responses to B and rare responses to C, was inconclusive, \( BF = .53, t(24) = 1.45, p = .16 \). The proportion of common responses to the A stimulus was significantly greater than .5, \( t(24) = 6.14, p < .001 \). The Bayesian evidence for the presence or absence of a difference from .5 for the DE stimulus (shared cue test) was inconclusive, \( BF = .51, t(24) = 1.41, p = .17 \). The Bayesian evidence for the presence or absence of a difference in the proportion of common responses to the BC and DE stimuli was also inconclusive, \( BF = .62, t(24) = 1.57, p = .13 \). Response times for participants in the test phase did not significantly differ between B and C trials, \( t(24) = 1.24, p = .23 \), or between D and E trials, \( t(24) = .62, p = .54 \).
4.2. EXPERIMENT 8

Imaging Analyses

**C - B activation comparison**

The C-B comparison aims to capture the difference in activations reflective of the reduction of prediction error presumed to drive the IBRE. There should be greater activation for the C stimulus compared to the B stimulus, due to the greater occurrence of prediction error for the C stimulus. The ROI analysis (thresholds of $p < .005$ and 28 contiguous voxels) revealed a number of brain regions that exhibited greater activations for C (rare stimulus) and B (common stimulus) test trials (Figure 4.4). When comparing responses for the C stimulus to responses for the B stimulus, greater activation was noted in the right Medial Frontal Gyrus (cluster size = 226, peak voxel $x = 3, y = 55, z = 17$, BA’s 9,10 and 32), the bilateral caudate body (left cluster size = 104, peak voxel $x = -12, y = 13, z = 12$, right cluster size = 167, peak voxel $x = 10, y = 3, z = 17$), the left Anterior Cingulate (cluster size = 40, peak voxel $x = -3, y = 43, z = 8$) and the right Middle Frontal Gyrus (cluster size = 37, peak voxel $x = 34, y = 6, z = 41$).

The activated areas for the whole-brain analysis can be seen in Table 4.3. Outside of brain areas already identified in the ROI analysis the left cerebellum was activated (cluster size = 125, peak voxel $x = -25, y = -70, z = -28$). This is perhaps unsurprising given its extensive implication in a wide range of cognitive tasks including learning (Desmond & Fiez, 1998), and a greater activation for C stimuli is perhaps reflective of the fact that learning is still ongoing for this stimulus.

**E - D activation comparison**

The E-D comparison is similar to the previous comparison, however it serves a different purpose. E and D serve as frequency matched controls for C and B, so if the differences in C-B are a result of the differing frequency, similar differences should be found for the
Table 4.3: Brain regions activated for whole brain analysis of C-B. The thresholds used were $p < .001$ and 106 contiguous voxels.

<table>
<thead>
<tr>
<th>Region</th>
<th>Cluster size</th>
<th>BA</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right Medial Frontal Gyrus</td>
<td>229</td>
<td>10</td>
<td>3</td>
<td>55</td>
<td>17</td>
<td>4.38</td>
</tr>
<tr>
<td>Right Anterior Cingulate</td>
<td>32</td>
<td>3</td>
<td>3</td>
<td>39</td>
<td>15</td>
<td>3.88</td>
</tr>
<tr>
<td>Right Superior Frontal Gyrus</td>
<td>9</td>
<td>12</td>
<td>5</td>
<td>55</td>
<td>20</td>
<td>3.75</td>
</tr>
<tr>
<td>Left Cerebellum</td>
<td>125</td>
<td></td>
<td>-25</td>
<td>-70</td>
<td>-28</td>
<td>3.94</td>
</tr>
<tr>
<td>Left Cerebellum</td>
<td></td>
<td></td>
<td>-29</td>
<td>-66</td>
<td>-33</td>
<td>3.30</td>
</tr>
<tr>
<td>Right Caudate Body</td>
<td>189</td>
<td></td>
<td>6</td>
<td>-1</td>
<td>13</td>
<td>3.87</td>
</tr>
<tr>
<td>Right Anterior Nucleus</td>
<td></td>
<td></td>
<td>10</td>
<td>-13</td>
<td>18</td>
<td>3.78</td>
</tr>
<tr>
<td>Right Thalamus</td>
<td></td>
<td></td>
<td>8</td>
<td>-6</td>
<td>6</td>
<td>3.70</td>
</tr>
</tbody>
</table>

E-D comparison. However, the EXIT model and previous studies (Kruschke, 2001a; Wills et al., 2014) suggest a differing response pattern for E and D compared to C and B. If these predictions hold true, then by extension the areas that exhibit greater activation in the C-B comparison should also differ in the E-D comparison.

A ROI analysis examined activations for the E stimulus compared to the D stimulus and failed to find any areas that showed a significant difference in activation. Although this is unsurprising theoretically, due to the lack of a shared cue, these analyses were conducted to both stay consistent with the other comparisons and to characterise this comparison given its use in further analysis. Whole brain analysis also failed to show any areas with a significant difference in activation.

**(C - B) - (E - D) comparison**

![Brain images](image)

$\begin{align*}
&x = 13, y = 12, z = 10 \\
\end{align*}$

*Figure 4.5: Areas that show greater activation for the C-B comparison compared to the E-D comparison under ROI analysis. The thresholds used were $p < .005$ and 28 contiguous voxels.*

This comparison is the critical analysis for this experiment. The previous comparisons differ in one key way; the presence or absence of a shared cue when training with those stimuli.
4.2. EXPERIMENT 8

Whilst any difference in the areas of the brain activated between these comparisons can be attributed to this factor, the (C - B) - (E - D) comparison provides a direct test of this difference.

ROI analysis revealed a number of brain regions exhibiting greater activation for the C-B comparison compared to the E-D comparison. Greater activation was noted in the bilateral caudate body (left cluster size = 58, peak voxel $x = -14, y = 8, z = 19$, right cluster size = 178, peak voxel $x = 17, y = 14, z = 11$). Several clusters were noted in the bilateral Middle Frontal Gyrus (left cluster size = 31, peak voxel $x = -40, y = 45, z = 8$, first right cluster size = 39, peak voxel $x = 51, y = 16, z = 30$, second right cluster size = 39, peak voxel $x = 38, y = 41, z = 23$, BA’s 8, 9, 10 and 46) and the right Medial Frontal Gyrus (cluster size = 28, peak voxel $x = 14, y = 48, z = 11$).

The activated areas for the whole-brain analysis can be seen in Table 4.4. Outside of the regions already identified for the ROI analysis, two clusters were identified in the right Superior Temporal Gyrus (cluster size = 119, peak voxel $x = 45, y = -44, z = 13$) and the left Middle Temporal Gyrus (cluster size = 120, peak voxel $x = -60, y = -47, z = 2$). These areas have been linked to cognitive memory functions (Chen, Kuo, Chiang, Tseng, & Lin, 2013; Yonelinas, Hopfinger, Buonocore, Kroll, & Baynes, 2001). The greater activation noted in these areas for the C-B comparison relative to the E-D comparison suggests that learning the C and B outcomes is more cognitively demanding than learning the E and D outcomes. This is perhaps due to presence of the shared cue, A, on the AB and AC training trials and its imperfect nature as a predictor.

<table>
<thead>
<tr>
<th>Region</th>
<th>Cluster size</th>
<th>BA</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
<th>$z-score$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right Superior Temporal Gyrus</td>
<td>119</td>
<td>13</td>
<td>45</td>
<td>-44</td>
<td>13</td>
<td>3.85</td>
</tr>
<tr>
<td>Right Superior Temporal Gyrus</td>
<td>22</td>
<td>52</td>
<td>-60</td>
<td>-47</td>
<td>12</td>
<td>3.84</td>
</tr>
<tr>
<td>Right Superior Temporal Gyrus</td>
<td>39</td>
<td>41</td>
<td>-60</td>
<td>-47</td>
<td>6</td>
<td>3.78</td>
</tr>
<tr>
<td>Left Middle Temporal Gyrus</td>
<td>120</td>
<td>22</td>
<td>-60</td>
<td>-47</td>
<td>2</td>
<td>3.79</td>
</tr>
<tr>
<td>Left Inferior Temporal Gyrus</td>
<td>37</td>
<td>-51</td>
<td>-65</td>
<td>-5</td>
<td>5</td>
<td>3.76</td>
</tr>
<tr>
<td>Left Inferior Temporal Gyrus</td>
<td>37</td>
<td>-55</td>
<td>-54</td>
<td>0</td>
<td>0</td>
<td>3.58</td>
</tr>
</tbody>
</table>

Table 4.4: Brain regions activated for the whole brain analysis of the comparison of the C-B comparison and the E-D comparison. The thresholds used were $p < .001$ and 106 contiguous voxels.
4.2. EXPERIMENT 8

Discussion

Experiment 8 found preferentially rare responding to the BC cue. This represents a significant IBRE, without the flaw present in the analysis of O’Bryan et al. (2018). Unlike Wills et al. (2014) no significant B>C effect was found. Although this procedure is adapted from that of Experiment 3 in Chapter 2, which found the effect, it is unsurprising that the present experiment doesn’t find it, given the unreliability of finding the effect in previous IBRE work. It is possible that participants may have been more motivated to learn, due to the isolated nature of the fMRI procedure leading to the lack of the effect. The lack of the effect could also be due to the increased size of the stimuli in this study, in terms of the visual angle subtended. The evidence for a difference between the BC and DE cues was inconclusive behaviourally.

The ROI analyses identify a number of specific brain regions linked to prediction error correlate with the IBRE, including the caudate body, the middle frontal gyrus, the medial frontal gyrus and the anterior cingulate cortex. Further, these areas show greater activation for the C cue relative to the B cue. This follows the predictions and results of Wills et al. (2014) and suggests an attentional advantage for the C cue. The analyses do not identify any areas that show a difference in activation for the E cue relative to the D cue. The critical analysis identifies several areas that show a significantly greater activation difference between the C and B cues; relative to the activation difference between the E and D cues. These results provide strong support for the error-driven learning account of Kruschke (2001b).

The areas identified in the ROI analysis have all been previously linked to prediction error; but also have links to other cognitive processes. The caudate body has been linked to associative learning (Seger & Cincotta, 2005), so it is unsurprising that it is linked to the occurrence of the IBRE, a learning phenomenon. Attentional allocation has been linked to the anterior cingulate cortex (Pardo, Pardo, Janer, & Raichle, 1990). This is again unsurprising; the EXIT model Kruschke (2001b) implements attentional allocation in its mechanisms of attentional shift. Finally, both the middle frontal gyrus and the medial frontal gyrus have been linked to a number of executive functions, including attention and working memory (Lebedev, Messinger, Kralik, & Wise, 2004; Miller & Cohen, 2001). While the results of this experiment support the link between prediction error and the IBRE, future neuroscientific investigation could explore the extent to which these other processes
are also linked to the IBRE.

It is worth noting that a number of areas were identified outside of the area used for the ROI analyses. One area in the left cerebellum showed a difference in activation for the C cue relative to the B cue. This area has been linked to learning and so it is perhaps unsurprising that it shows greater activation, given learning is more likely to still be ongoing for the C cue compared to the B cue. Two areas that showed a difference in activation in the critical analysis were in the left and right temporal gyri. These areas have been linked to cognitive memory functions and so suggest that the outcomes linked to the C and B cues take more cognitive effort to learn.

The predictions and results of Wills et al. (2014) are supported by the results of this experiment. Interestingly, none of the brain regions reported as showing activation differences in O’Bryan et al. (2018) showed any difference in activation in the whole-brain analyses in this experiment. This is potentially due to the fact that the present experiment reports a significant IBRE, and the flaw within the analysis method O’Bryan et al. (2018) suggests that they do not. As such, the activation differences noted in the regions in O’Bryan et al. (2018) are most likely linked to the representation of visual object categories, given their use of three different visual stimuli types and previous work linking the areas to this Grill-Spector and Weiner (2014).

Chapter Discussion

The experiment in this chapter had three aims; to produce a significant IBRE in an fMRI methodology, to identify activations correlated with the IBRE in brain areas and to evaluate the differences between the C and B cues, the matched frequency E and D cues and the difference of these differences. It succeeds in all three of these aims, providing the first fMRI study of the IBRE (O’Bryan et al. (2018) did not convincingly demonstrate the presence of an IBRE).

Experiment 8 reports a significant IBRE in an fMRI methodology, a result that is behaviourally consistent with previous work Wills et al. (2014). However, it does not find a significant $B > C$ effect, with the proportion of common responding to B being ordinally greater than rare responding to C but with inconclusive Bayes. This is potentially due to the isolated nature of the fMRI methodology; driving participants to be more motivated to learn. The behavioural results of Experiment 8 are not as strong as those of Wills et al. (2014) due to this. Further, the behavioural analysis does not find conclusive evidence
of a difference in responding between the BC and DE cues. Previous work has found this difference (Kruschke, 2001a; Medin & Edelson, 1988; Wills et al., 2014), suggesting that the shared cue is required to produce the IBRE, but the present experiment cannot conclude if it is required or not.

It is worth noting that the conservative nature of the fMRI analysis method in this chapter does cause limitations in what can be concluded from the data. As this experiment aimed to be the first to find the IBRE significantly, and identify brain regions linked to the effect, an ROI analysis was used. However, this approach does not take into account individual differences in these areas between participants. Further, it cannot identify the development of these activations across time. It is possible that activation spreads across a network of brain areas within the regions of the ROI across time, and this analysis method is not sensitive enough to detect this. Future work should try to address these issues, using a methodology that is accurate on both a spatial and temporal level.

The areas identified in the ROI analysis were drawn from neuroscience studies investigating prediction error. Prediction error is an integral part of the error-driven attentional account advocated by Kruschke (2001b); which suggests that the reduction of prediction error drives learning. As Experiment 8 identified activation difference in specific regions within the areas linked to prediction error, within an experiment that shows a significant IBRE, it provides support for the error-driven attentional account. Further evidence for this account comes from the results of the fMRI analysis; primarily the fact that there was a greater difference in activations between the C and B cues compared to the D and E cues.
Chapter 5

General Discussion

General Findings

In this thesis I have provided several datasets that help to further our understanding of the IBRE. Chapter 2 identified a procedure capable of finding the IBRE and the B>C effect in conjunction, something previously found in only a handful of studies with differing procedures (Bohil et al., 2005; Wills et al., 2014; Winman et al., 2005). The chapter also contains the first investigation of the effects of context on an IBRE procedure. Finally, the response patterns for the novel cue in Experiment 4 within Chapter 2, where responding is preferentially common, is inconsistent with the response patterns reported in previous work (Johansen et al., 2007; Juslin et al., 2001). This pattern of results is discussed in more detail later in this chapter. Chapter 3 provides statistical analysis on the effect of extended training on the IBRE and responding to a novel cue, and attempts to replicate Juslin et al. (2001)’s findings. Whilst their suggestion that extended training would reduce the magnitude of the IBRE seems to be incorrect, their prediction for a reduction in rare outcome responses to the novel cue with extended training was found within all three experiments in this chapter. Chapter 4 provides the first fMRI procedure to successfully find the IBRE and identifies a number of areas within the brain where increased activation correlates with the effect; many of which have previous links to prediction error. The implications of these findings are discussed below.

The Context Explanation

Le Pelley et al. (2016) offered a potential reconciliation for the occurrence of the IBRE and the B>C effect in conjunction, appealing to the role of experimental context. The detailing of this explanation culminated in a set of testable predictions about how the response patterns for the BC, B and C and N cues would change if the experimental context were to change. These predictions suggest that preferentially rare responding to BC would increase, that rare responses to C would become greater than common responses to
5.2. THE CONTEXT EXPLANATION

B and that *common* responding to the N cue would shift to random responding. The experiments within Chapter 2 represent the first attempt to test these predictions formally. In short, these predictions were found to be incorrect.

Experiment 1 found the IBRE in conjunction with a response pattern of C>B responding in both same context and different context conditions. These results are explainable if one assumes that for this experiment, context was ignored by the participants. However, this does not explain the response pattern to the novel cue. Whilst responses to N were statistically at random in both conditions, there was evidence of greater *common* responding in the same-context condition relative to the different-context condition. The paradox of these response patterns co-occurring is something the context explanation cannot accommodate. Experiment 2 found random responding to the IBRE test cue, a B>C effect and random responding to the N cue across both same and different context conditions. The criterion set in this experiment was more lax in comparison to Experiment 1 and it seems that participants did not learn much as a result. The data produced by Experiment 2 is therefore not particularly diagnostic of any account. Experiment 3 found both the B>C effect and inconclusive evidence for the IBRE, without the presence of a context manipulation. Experiment 4 found no change in the preferentially rare responses to BC with a shift in context, and that *common* responses to B remained significantly greater than rare response to C even when the context was changed. The prediction for the novel cue N was found ordinally, but Bayes analysis found that the evidence for this change was inconclusive.

A criticism for these experiments worth acknowledging is that the change in context was not salient enough to elicit change in participants’ behaviour. This criticism doesn’t hold for Experiment 1, with noted changes in responding for the A and N cues between same and different context conditions. Although the response pattern for the N cue in Experiment 4 does seem to suggest some effect of context, the strongest evidence against this criticism remains the change in the novel cue response pattern in Experiment 1. The novel cue response pattern of Experiment 1 will be discussed more thoroughly later in this chapter.

In summary, the context explanation alone seems inadequate for explaining the IBRE and B>C effects. However, one final point worth considering relevant to the paradoxical results of Experiment 1 lies in the different abstract nature of the novel cue in comparison to the other cues. The presentation of the compounds (BC) and individual cues (B,C) at
test is novel, but the cues themselves are presented during training. The novel cue is the only cue that is entirely new and so is abstractly different from the other cues. Due to this, it may still be the case that learning about the experimental context is important, but that this knowledge is only relevant in specific cases i.e. for making a judgement about a completely novel cue.

**Evaluation of EXIT**

Kruschke (2001a) proposed EXIT as a formal model implementing the ideas of error driven attentional theory. EXIT represents the best current explanation of the IBRE, capable of capturing the IBRE and the effects that occur in conjunction with it (Kruschke, 2001a, 2003, 2005b). The pattern of responses for a novel cue are something that no study has tasked EXIT with simulating, likely due to the scarcity of studies employing a novel cue in their procedure (Don & Livesey, 2017; Johansen et al., 2007; Juslin et al., 2001). As a result of the above, simulations of EXIT were conducted fitting the model to the behavioural results of Experiments 1, 2 and 4 of Chapter 2 and Experiments 5, 6 and 7 of Chapter 3.

The simulations within Chapter 2 demonstrate the capabilities of EXIT at capturing IBRE results. The key results in these experiments were the BC, B and C, and N cue response patterns. EXIT was able to capture the key results of Experiment 1, the IBRE, C>B and common responding shifting to random responding for the novel cue between context conditions. The simulations for Experiments 2 and 4 demonstrate a known issue with EXIT; that the model can only capture B>C under specific conditions. EXIT was able to capture the other key results for these experiments; just not the B>C response patterns. The specific fix for this implemented in Kruschke (2003) was not used in these simulations; due to concerns that this might harm the model's ability to capture the novel cue response patterns. Future simulation work using these datasets intends to determine if the best fitting parameters change if the B>C response pattern is more heavily weighted.

The simulations within Chapter 3 highlight a key issue within the EXIT model. The key results in these experiments were the BC, B and C, N and the blank cue response patterns. EXIT was able to capture the response patterns for the BC, B and C, and the blank cue effectively; but not the response pattern for the N cue. The response pattern for the novel cue N in these experiments is that responding starts out preferentially rare at low levels of training, and then shifts towards random responding at higher levels of training. At no point in any of the simulations for the experiments within this chapter was EXIT able to
capture this low training level response pattern. Considering the mechanics of the model's formal implementation, EXIT has no mechanism that could cause this response pattern to arise. As such, without some form of major change to the model's architecture, EXIT will continue to be unable to capture these results.

Wills and Pothos (2012) outline an argument containing a number of specific concepts for effective model comparison. One of the concepts is that models’ successes should be irreversible and that arbitrarily-variable parameters should be avoided. This means that a model should not have a specific set of parameters for individual experiments, instead possessing a set of general parameters to capture phenomena across experiments. Whilst this thesis does individually fit experiments, this represents a first point of investigation for the ability of EXIT to capture these results. Future simulation work will address this issue, tasking EXIT with fitting all of the experiments at once, to determine the best fitting parameters and to evaluate the adequacy of this fit.

**Neural Correlates of the IBRE**

Whilst some neuroscience methodology has been used to investigate the IBRE (Wills et al., 2014), no published study currently makes use of fMRI to evaluate the specific brain areas linked to the IBRE. O’Bryan et al. (2018) attempted to investigate this; however a flaw within their analysis method casts doubt on the validity of their findings. Chapter 4 reports the first fMRI study to produce a significant IBRE.

The fMRI ROI analysis supported the predictions made in Experiment 8; that there would be a greater brain activation for the C relative to the B cue, that there would be no difference in brain activation between the E and D cues and that there would be a difference between the two previous differences. The specific brain areas identified through the ROI analysis were those that were previously linked to prediction error (Botvinick et al., 2001; Critchley et al., 2005; Fletcher et al., 2001; Fouragnon et al., 2018; Garrison et al., 2013; Hoffman & Falkenstein, 2011; McClure et al., 2003; O’Doherty et al., 2003; Pagnoni et al., 2010; Park et al., 2010; Rodriguez, 2009; Spoormaker et al., 2011; Turner et al., 2004).

These results provide further support for the error-driven attentional account of the IBRE. Prediction error is assumed to be a key component of this explanation; specifically, the account suggests that learning occurs through a necessity to reduce the further occurrence of prediction error, and attentional shift facilitates this. The greater activation noted for the C compared to the B cue suggests that the C cue is more strongly linked to prediction error.
occurrence; perhaps unsurprising given its less frequent presentation. This would suggest that more attention is paid to C when the cues occur in compound, in order to minimise prediction error. The lack of a difference in activation between E and D cues serves to further support the idea that the shared cue A present during training is required for the effect to occur (Kruschke, 2001a; Medin & Robbins, 1971; Wills et al., 2014). The critical difference of differences analysis does support the idea that the shared cue is required for the effect, but this evidence is hampered by the behavioural results. Specifically, the evidence for an effect of shared cue on responding is inconclusive. Therefore, the most that can be said for this analysis is that attentional reallocation is more pronounced for cues that are presented in compound with a shared cue during training.

**Further Directions**

Whilst this thesis contains a more thorough investigation of the novel cue and how responding to it changes, further investigation would be beneficial. Consider the response patterns to a novel cue within this thesis as a whole. The three experiments within Chapter 3 all demonstrate rare responding to the novel cue that shifts to random responding as training length increases all within the same experimental context. Experiment 1 of Chapter 2 is inconsistent with this if we consider the same-context condition. In this condition, whilst responding to the novel cue is statistically at random, common responding is significantly greater than in the different-context condition. One potential explanation for this is that the length of training differs between the experiments. Participants in Experiment 1 experience generally more training trials than participants in Experiments 5 and 6. However, participants in the high level training condition of Experiment 7 receive roughly twice as much training as those of Experiment 1 and yet exhibit random rather than preferentially common responding; so the results remain inconsistent. Another considerable difference between the training procedures used in these experiments is based on how participants were trained. Experiment 1 of Chapter 2 uses criterion-base training, where participants only proceed to test after achieving a certain level of accuracy, whilst the experiments in Chapter 3 all have fixed training lengths. The effect of criterion-base training on learning and responding to a novel cue within the IBRE procedure is something that merits further investigation, to determine the specific circumstances under which responding to a novel cue is either preferentially common or preferentially rare.

EXIT represents the formal model best capable of capturing the IBRE and surrounding
effects. This thesis does not dispute that. However, as demonstrated by the experiments within Chapter 3, there exists a series of response patterns around the novel cue that the model cannot capture. Therefore what follows is an informal consideration of how EXIT may be developed to capture these results. Considering the work of Juslin et al. (2001), the response pattern for the novel cue arises more naturally from an account that allows for learning about the concept of novelty. Given that EXIT arose as a development to allow for learning about attention itself, further developing the model to allow learning about novelty seems appropriate. So, how might this be achieved? The Attention To Rules and Instances in a Unified Model (ATRIUM) offers some insight into this. ATRIUM (Erickson & Kruschke, 1998; Kruschke & Erickson, 1994) builds on the ALCOVE model in a different way to EXIT, utilising both the exemplar system of ALCOVE and a system representing a number of simple rules. The rules ATRIUM uses were developed for other category learning tasks, but this is not necessarily an issue. Creating a rule-based system within EXIT that is able to learn simple rules about concepts such as novelty could allow for the capture of the results within Chapter 3. An example rule would be: if presented with a novel cue, respond with the rare outcome.

The mechanism for how this rule-based system would control responding is up for debate; whilst the two systems of ATRIUM control over responding is adjusted by a competitive gating system, it would perhaps be better for a rule-based system within EXIT to only influence responding on certain trials. It would seem sensible for such a system to have a high amount of influence early in learning on trials such as a novel cue; where nothing is known about the cue. Early in learning the system would have learnt that more novel, less frequent cues lead to novel, less frequent outcomes i.e. AC leading to the rare outcome. The system would then influence the model to produce a rare outcome response. If the mechanism that mediates the influence of the rule-based system assigns less control to the system as more is learnt, the model would generate responses to a novel cue based on the bias cue already implemented within EXIT. This could then produce the preferentially common responding to the novel cue noted in Experiment 1. Whilst this possible path for development is currently informal, further debate and discussion goes beyond the scope of this thesis and so is left for future work.
Final Remarks

As the era of information continues, one of the key paths for advancement as a civilisation is mechanisation. Machines are beginning to be used for jobs typically performed by people, with machine learning and artificial intelligence development being pushed by companies such as Tesla with their forays into self-driving cars (Tesla, 2016). One of the key points for developing artificial intelligence for use in our world is that they are able to approximate human behaviour. This requires us to understand the processes of human intelligence; such as learning. Artificial Intelligences able to play strategic board games such as Chess and Go have already been developed, as well as more advanced artificial intelligences capable of playing complex strategy video games such as Defense of the Ancients 2 (OpenAI, 2018). The level of complexity of such a task starts to approach the continuous stream of information present in our world's environment. One thing worth considering is the way in which these systems achieve this. They replicate rational human behaviours, based around actions that are directly beneficial to the goal of winning. One of the major issues for the systems detailed in OpenAI (2018) was noted when having them directly compete against human players. Specifically, the systems had no way to respond when the human players performed irrational actions, enabling the human players to win by acting in a non-optimal manner. Future development of AI systems should aim to capture both rational and irrational human behaviours, adapting over time to learn where these behaviours are most useful. Researching non-rational behavioural phenomena such as the IBRE is essential to this goal, and formal models such as EXIT provide a starting point for how AI may begin to capture such behaviours.
References


REFERENCES


person impressions. *Journal of Experimental Social Psychology*, 68, 146 - 156.


