

# Neural Correlates of the Inverse Base Rate Effect

Angus B. Inkster<sup>1</sup>, Fraser Milton<sup>2</sup>, Charlotte E. R. Edmunds<sup>3</sup>, Abdelmalek Benattayallah<sup>1</sup> and Andy J. Wills<sup>1</sup>

<sup>1</sup>Brain Research and Imaging Centre, University of Plymouth

<sup>2</sup>University of Exeter

<sup>3</sup>Queen Mary, University of London

## Author Note

This project was funded by a full Ph.D. scholarship from the University of Plymouth to the first author. The authors wish to thank Anna Robertson and Gemma Williams for their help in preparing some of the experimental materials, as part of an undergraduate research placement scheme at the University of Plymouth.

The raw imaging and behavioral data, as well as the analysis and modeling scripts for the experiment within this paper are available at <https://osf.io/yw6fj/>.

Participants for this study were recruited with no specific exclusion on the basis of age, sex or race.

Correspondence concerning this article should be addressed to Angus Inkster, School of Psychology, University of Plymouth, Drake Circus, Plymouth, PL4 8AA. E-mail: [angus.inkster@plymouth.ac.uk](mailto:angus.inkster@plymouth.ac.uk)

## 1 Abstract

2 The Inverse Base Rate effect (IBRE; Medin & Edelson, 1988) is a non-rational behavioral phenomenon in  
3 predictive learning. Canonically, participants learn that the AB stimulus compound leads to one outcome and that  
4 AC leads to another outcome, with AB being presented three times as often as AC. When subsequently presented  
5 with BC, the outcome associated with AC is preferentially selected, in opposition to the underlying base rates  
6 of the outcomes. An error-driven learning account (Kruschke, 2001b) is the leading current explanation of the  
7 IBRE. A key component of this account is prediction error, a concept previously linked to a number of brain  
8 areas including the anterior cingulate, the striatum and the dorsolateral prefrontal cortex. The present work is the  
9 first fMRI study to directly examine the IBRE. Activations were noted in brain areas linked to prediction error,  
10 including the caudate body, the anterior cingulate, the ventromedial prefrontal cortex and the right dorsolateral  
11 prefrontal cortex. Analysing the difference in activations for singular key stimuli (B and C), as well as frequency  
12 matched controls, supports the predictions made by the error-driven learning account.

13 **Keywords:** cognitive neuroscience, human learning, fMRI, prediction error, inverse base rate effect

## 14 1 Introduction

15 Learning is a process that enables the use of past and present information to adapt to and overcome present  
16 and future challenges. The amount of environmental information present on a moment-to-moment basis is  
17 large, and so humans have evolved to prioritize the most relevant information. However, the same processes of  
18 prioritization can sometimes lead to irrational decisions. The Inverse Base Rate Effect (IBRE; Kruschke, 1996,  
19 2001a; Medin & Edelson, 1988; Shanks, 1992) is one example of an irrational decision-making behavior that  
20 seems to occur in this way.

21 In its canonical form, shown in Table 1, the IBRE involves participants being trained under a simulated  
22 medical diagnosis procedure. They are presented with a patient with one of two different pairs of symptoms,  
23 and asked to make a judgment, diagnosing the patient with one of two fictitious diseases. For the purposes  
24 of this example, we refer to them as “Jominy Fever” and “Phipps Syndrome”. Participants see patients for  
25 whom the correct diagnosis is “Jominy Fever” three times as often as those for whom the correct diagnosis is  
26 “Phipps Syndrome”. “Jominy Fever” is therefore referred to as the common disease, because its base rate is  
27 higher. “Phipps Syndrome” is referred to as the rare disease, due to its lower base rate. The symptom pairs can  
28 be considered abstractly as AB and AC. So, a participant might be presented with a patient suffering from “ear  
29 aches” and “skin rash” (AB) where the correct diagnosis is “Jominy Fever” (common). They then might see a  
30 patient suffering from “ear aches” and “back pain” (AC), with the correct diagnosis being “Phipps Syndrome”  
31 (rare). In this example “skin rash” (B) is perfectly predictive of “Jominy Fever” (common), while “back pain”

**Table 1.** Canonical IBRE experimental design.

Training trials	
(relative frequency)	Test trials
$AB \rightarrow \textit{common}$ (x3)	$BC \rightarrow \textit{rare}$
$AC \rightarrow \textit{rare}$ (x1)	

32  
33 (C) is perfectly predictive of “Phipps Syndrome” (rare). The symptom “ear aches” (A) is uninformative. After  
34 being trained in this manner, participants are then presented with both perfectly predictive symptoms, “skin  
35 rash” (B) and “back pain” (C). If participants make use of the base rate of the two diseases, they should make  
36 the rational diagnosis of the more common disease, “Jominy Fever”. However, the majority of participants  
37 preferentially diagnose the patient with the rarer disease, “Phipps Syndrome”. This pattern of responding is  
38 called the IBRE.

39 [Currently, the best explanation of the IBRE](#) is the error-driven learning account implemented within the  
40 EXemplar-based attention to distinctive InpuT (EXIT) formal model (Kruschke, 2001b). Kruschke’s error-driven  
41 learning account suggests that, during learning, participants endeavor to reduce the number of errors they make  
42 through the shifting of attention. This account predicts that, due to the more frequent occurrence of AB compared  
43 to AC, participants learn more about A and B than about C. When they encounter AC, participants initially  
44 respond with the common disease due to the presence of A, leading to a prediction error. Attention then shifts  
45 away from A and towards C when presented with A and C together, in order to promote new learning and reduce  
46 the further occurrence of prediction errors.

47 EXIT assumes that this attentional reallocation, driven by prediction error, is persistent. As a result, when  
48 presented with B and C together during test, the attention to C is greater than the attention to B, resulting in  
49 the IBRE. The EXIT model’s assumption of the persistence of attentional reallocation is supported by greater  
50 eye-tracking dwell time for C compared to B when presented with BC at test (Kruschke, Kappenman, & Hetrick,  
51 2005). Attention also persists to singly-presented cues at test, as demonstrated electrophysiologically by a  
52 selection negativity/positivity for C over B when each cue is presented alone on separate test trials (Wills, Lavric,  
53 Hemmings, & Surrey, 2014). This attentional persistence to singly-presented cues at test is also predicted by  
54 EXIT, and is the central prediction investigated in the current study.

55 One strength of EXIT’s error-driven learning account is that it explains not only the IBRE but also other  
56 concurrent response patterns that often occur. When presented with the A cue alone, responding is preferentially  
57 common, following the base rate of the two diseases. This is explained by assuming that participants learn to  
58 associate A with the common disease more than the rare disease. Another phenomenon occurs when participants

59 are also trained with control cues for B and C, labeled as D and E. These cues are matched for frequency but  
60 lack a shared cue (A) during training. This shared-cue effect is characterized by the IBRE disappearing for  
61 the control stimuli, i.e. participants do not respond preferentially rare when presented with DE. This has been  
62 found in a number of studies (e.g. Kruschke, 2001a; Medin & Edelson, 1988). The error-driven learning account  
63 predicts this effect because, in the absence of a shared cue, there is nothing to cause attentional reallocation on  
64 the rare-outcome trials. While alternative accounts of the IBRE, such as the relative novelty account (Binder &  
65 Estes, 1966), and the eliminative inference account (Juslin, Wennerholm, & Winman, 2001) can accommodate  
66 the basic IBRE, they fail to account for the shared-cue effect (Kruschke, 2001a; Wills et al., 2014).

67 The only previous published fMRI study of the IBRE was conducted by O'Bryan, Worthy, Livesey, and  
68 Davis (2018). They made use of an atypical IBRE procedure involving real-world visual categories (scenes,  
69 faces and objects) as stimulus features to allow their use of multi-voxel pattern analysis. While this approach  
70 was well motivated, one consequence of this atypical procedure was the lack of a compelling behavioral IBRE in  
71 their study. Specifically, the defining feature of the IBRE is the presence of greater rare than common responses  
72 to BC. O'Bryan et al. report the presence of a numerical effect in that direction, without reporting inferential  
73 statistics for this contrast; our analysis of their raw data indicates Bayesian evidence for the absence of the IBRE  
74 in their study,  $BF_{10} = .27$ . The inferential tests reported by O'Bryan et al. provide evidence for base-rate neglect  
75 rather than the IBRE.<sup>1</sup>

76 In the current study, we employed a more standard procedure from our previous work, known to robustly  
77 demonstrate the IBRE (Inkster, 2019; Wills et al., 2014). We had two predictions, based on EXIT, the leading  
78 account of the IBRE, and on our previous electrophysiological work (Wills et al., 2014). Our first prediction,  
79 well supported in general terms by previous neuroimaging work on the correlates of prediction error, was that  
80 the striatum, the medial anterior prefrontal cortex, and the anterior cingulate would show more activation for AC  
81 than for AB during training. This is because AC results in more prediction errors than AB behaviorally, and  
82 because previous work, including two major meta-analyses (Fouragnan, Retzler, & Philiastides, 2018; Garrison,  
83 Erdeniz, & Done, 2013), implicate these areas in the processing of prediction errors. There is also good evidence  
84 that the right dorsolateral prefrontal cortex is involved in the processing of prediction errors (Fletcher et al., 2001;  
85 Fouragnan et al., 2018; Turner et al., 2004). We thus defined a region of interest (ROI) for all of our analyses  
86 that comprised these four areas.

87 As discussed by Fouragnan et al. (2018), the activity in brain areas associated with prediction error is  
88 likely due to a number of different processes, including outcome valence processing, attentional processing –

---

<sup>1</sup>O'Bryan et al. (2018) report that the proportion of rare responding to BC (.5) is significantly greater than the base-rate of .25. This supports the presence of base-rate neglect, but lacks the greater rare compared to common responding indicative of an IBRE. Similarly, their demonstration of significantly greater rare responding to BC compared to rare responding to A suggests base-rate neglect is smaller for A than BC, but does not show the presence of an IBRE.

89 sometimes described as “surprise” processing or the modulation of associability (Mackintosh, 1975; Pearce &  
90 Hall, 1980) – as well as the calculation of signed prediction error that is most commonly associated with the  
91 term *prediction error* (and as instantiated by, for example, the Rescorla-Wagner (Rescorla & Wagner, 1972) and  
92 temporal difference models (Sutton & Barto, 1987) .

93 Our second prediction for the current study concerns the possible attentional-processing role of prediction-  
94 error-associated brain areas, and comes from the EXIT model’s explanation of the IBRE. A key part of EXIT’s  
95 architecture is a back-propagation process, driven by prediction error, which adjusts future attention to stimuli  
96 in order to minimize errors. In the case of the IBRE procedure, when participants encounter AB, attentional  
97 changes are less frequent due to both A and B being associated with the common outcome and so there is less  
98 chance of an error being made (B because of it being a perfect predictor of the common outcome, A because it  
99 occurs more frequently with the common outcome than the rare). When they encounter AC, errors are dependent  
100 on the cue preferentially attended to and are more frequent, due to the disjoint of C being a perfect predictor of  
101 the rare outcome and A being associated more heavily with the common outcome. On these trials, EXIT predicts  
102 that when a prediction error occurs, cue attention for future AC trials is shifted such that more attention is paid  
103 to the C cue. The model assumes that these attentional changes are persistent, and that in order for the IBRE to  
104 occur this attentional reallocation persists into the test phase, producing the preferential rare outcome responding  
105 to BC at test (i.e. because C is attended more than B). Previous eye-tracking and neuroscience work (Kruschke  
106 et al., 2005; Wills, Lavric, Croft, & Hodgson, 2007; Wills et al., 2014) observed these persistent attentional  
107 changes, and other work (Fouragnan et al., 2018) acknowledge the possibility that other neuroscience studies of  
108 prediction error could be observing persistent attentional changes caused by prediction error; rather than (or as  
109 well as) the initial computation of prediction error.

110 In the context of the current study, our prediction is that this persistence of attentional reallocation would  
111 manifest as greater activation for cue C, presented alone at test, than for cue B, presented alone at test. Our  
112 assumption that attentional reallocation persists not only into the test phase, but also to singly-presented cues  
113 is supported by our previous neurophysiological work (Wills et al., 2014). Thus, our *a priori* prediction was  
114 that we would see greater activation for C than for B during test in our prediction-error ROI. If confirmed, this  
115 prediction would further support the EXIT account of the IBRE, and would suggest that the brain areas in which  
116 this difference was observed may be involved in the persistent attentional reallocation that can occur in response  
117 to prediction errors.

## 118 2 Methods

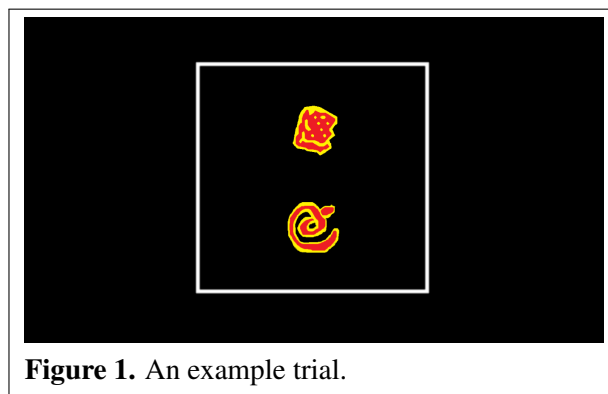
### 119 2.1 Participants

120 34 people were recruited from the University of Exeter participant pool. Participants received either course  
121 credit or £10. Participants gave informed consent according to procedures approved by the Psychology Ethics  
122 Committee, University of Exeter. Five participants' data were removed due to excessive head movements during  
123 the experiment, rendering their fMRI data unusable. Participants' accuracy in the final block of training was  
124 then assessed using a learning criterion. This criterion was identical to the one used in Wills et al. (2014),  
125 where participants scoring less than 72% in the final block of training were excluded from further analysis. This  
126 criterion represents the level of accuracy that cannot be attributed to random responding based on the block  
127 length of 18 trials. Applying this criterion necessitated the removal of 4 participants, resulting in a final data set  
128 of 25 participants.

### 129 2.2 Procedure

130 The abstract design and stimuli are identical to that of Wills et al.'s (2014) electrophysiological study, and  
131 can be seen in Table 2 and Figure 1 respectively. The stimuli are abstract shapes, referred to as "cells" due to the  
132 context of the experiment; a medical diagnosis task. The ratio of common to rare in this design (2:1) differs from  
133 the ratio in the canonical IBRE design (3:1). The reason for this is the same as in Wills et al.; it shortens study  
134 duration in order to avoid participant fatigue, given the necessarily long test phase required for a neuroscience  
135 study. Previous work (Inkster, 2019; Wills et al., 2014) has shown that a robust IBRE can be achieved with a 2:1  
136 ratio of common to rare.

137 In each phase of the experiment, trial order was randomized. Participants were asked to take on the role of  
138 a doctor, diagnosing patients with either "Jominy Fever" or "Phipps Syndrome" on the basis of the "cells" they  
139 were presented with. These instructions were given prior to them entering the scanner. The response key that



**Table 2.** Experimental design.

Training trials		Test trials	
(relative frequency)			
$A_1B_1 \rightarrow common$ (x2)		$A_1B_1, A_2B_2, A_3B_3,$	x4
$A_2B_2 \rightarrow common$ (x2)		$F_1D_1, F_2D_2, F_3D_3,$	
$A_3B_3 \rightarrow common$ (x2)		$A_1C_1, A_2C_2, A_3C_3,$	x2
$A_1C_1 \rightarrow rare$ (x1)		$G_1E_1, G_2E_2, G_3E_3,$	
$A_2C_2 \rightarrow rare$ (x1)		$B_1, B_2, B_3,$	
$A_3C_3 \rightarrow rare$ (x1)		$C_1, C_2, C_3,$	
$F_1D_1 \rightarrow common$ (x2)		$D_1, D_2, D_3,$	
$F_2D_2 \rightarrow common$ (x2)		$E_1, E_2, E_3,$	x 5
$F_3D_3 \rightarrow common$ (x2)		$A_1, A_2, A_3,$	
$G_1E_1 \rightarrow rare$ (x1)		$B_1C_1, B_2C_2,$	
$G_2E_2 \rightarrow rare$ (x1)		$B_3C_3, D_1E_1,$	
$G_3E_3 \rightarrow rare$ (x1)		$D_2E_2, D_3E_3$	

*Note.* Each abstract stimulus is represented by three “cells” randomized between participants. The subscripted numbers represent the specific “cell” tied to the abstract stimulus present on a trial. Example “cells” can be seen in Figure 1.

142 represented each disease was also explained to the participant before the task began, and was counterbalanced  
 143 between participants. The disease that was abstractly common or rare was also counterbalanced. The mapping  
 144 between cues and outcomes was deterministic e.g.  $A_1B_1$  was always followed by the common disease, and  $A_1C_1$   
 145 was always followed by the rare disease.

146 The experiment was displayed on a back-projection screen positioned at the foot end of the MRI scanner  
 147 and viewed via a mirror mounted on a head coil. Button-press responses and reaction times (RTs) were measured  
 148 using a fiber-optic button box. The training phase consisted of 10 blocks of 36 trials, making 360 trials in total.  
 149 Each trial began with a variable duration fixation cross presented in the center of the screen. The durations were  
 150 generated using an exponential distribution, following the method described in Haberg, Zito, Patria, and Sanes  
 151 (2001). The range of the durations was 250 ms - 3500 ms, with a mean duration of 1284 ms.

152 After the fixation cross, a grey view box was displayed on its own for 500 ms to indicate where the stimuli  
153 would appear. The “cell” stimuli appeared toward the top and bottom of the view box, with location randomized  
154 on each trial. The cells remained on screen for 2000 ms, during which time participants made their diagnosis  
155 using either the left or right button on the button box. After this, participants received corrective feedback for  
156 500 ms which included naming the correct diagnosis. If a response was not made within 2000 ms, participants  
157 instead received a time-out message.

158 Further instructions were given at the start of the test phase. Participants were informed that they would  
159 still diagnose patients and would see some cells that they had seen before, continuing to receive feedback for  
160 these cells. These were the same cue compounds presented during training, and were presented in the same  
161 ratio as in training. The first four rows of the test trials column in Table 2 represent these trials. Training  
162 trials for which participants received corrective feedback in the test phase are not always included in IBRE  
163 procedures, but this approach addresses the potential concern that performance will deteriorate over the course  
164 of the necessarily lengthy test phase, by providing additional learning in order to stabilize performance. This  
165 technique was employed successfully in both Wills et al. (2007) and Wills et al. (2014).

166 Participants were further told that they would see some cell combinations that they would not receive  
167 feedback for. These trials were novel to the test phase, and can be seen in the test trials column in Table 2 (row  
168 five onwards). The test phase consisted of 282 trials in total. The number of test trials was constrained such that  
169 the key test stimuli (B, C, D, E) were presented enough to adequately power the fMRI analyses, but that the test  
170 phase was not excessively long, so as to avoid participant fatigue.

171 The trial structure in the test phase was the same as in the training phase, but with the addition of single  
172 cells being presented in the center of the view box. The variable duration of the fixation cross had the same  
173 range of times as in the training phase, and a similar mean duration of 1226 ms.<sup>1</sup> On trials for which participants  
174 did not receive feedback, they instead received the message “DATA MISSING” and a series of question marks.

### 175 **2.3 Analysis of Behavioral Data**

176 Trials where participants timed out were removed from further analysis and constituted less than 1% of the  
177 total number of trials across all participants. In addition to conventional null-hypothesis tests, we also calculated  
178 Bayes Factors (*BF*) for theoretically-central analyses. These were calculated using the procedure recommended  
179 by Dienes (2011), implemented within an R script by Baguley and Kaye (2010). Predicted differences were  
180 estimated from a behavioral-only version of the same experiment previously run in our lab (Experiment 3;  
181 Inkster, 2019). As recommended by Dienes, we assumed a half-normal distribution for the prior with a mean of

---

<sup>1</sup>The slight difference in mean duration relative to the training phase results from discretizing the exponential distribution of times over a different, finite, number of trials.



182 zero and a standard deviation equal to the predicted difference. By convention, where  $BF > 3$ , the experiment  
183 has found evidence for the alternative hypothesis, whereas if  $BF < 1/3$ , the experiment finds evidence for  
184 the null hypothesis (Jeffreys, 1961). Values between a third and three are generally considered inconclusive,  
185 although they still carry information. For example, where  $BF = 2$ , this tells us that the experimental hypothesis  
186 is now about twice as likely as it was before we conducted the experiment.

## 187 **2.4 fMRI Data Acquisition**

188 Images were collected using a 1.5-T Gyroscan magnet equipped with a Sense coil (Philips, Amsterdam,  
189 The Netherlands). A T2\*-weighted echo-planar sequence was used (repetition time = 3000 ms, echo time =  
190 45 ms, flip angle =  $90^\circ$ , 32 transverse slices, field of view = 240 mm,  $3.5 \times 2.5 \times 2.5$  mm). The training phase  
191 comprised two runs of 242 scans, and the test phase two runs of 187 scans. Standard volumetric anatomical MRI  
192 was performed after functional scanning by using a 3-D T1-weighted pulse sequence (repetition time = 25 ms,  
193 echo time = 4.1 ms, flip angle =  $30^\circ$ , 160 axial slices,  $1.6 \times 0.9 \times 0.9$  mm).

## 194 **2.5 Analysis of fMRI Data**

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

202 In the individual participant models, the critical trials for comparisons (AB and AC for the training phase;  
203 B, C, D, E for the test phase) were included as individual regressors, with the other, non-critical, trial types and  
204 time-outs included as two further separate regressors of no interest. The duration of each event was modeled as  
205 the participant's RT for that trial, an approach advocated in Grinbrand, Erdeniz, Lindquist, Ferrera, and Hirsch  
206 (2008).

207 Our three principal analyses were conducted on comparisons of singly-presented cues in the test phase;  
208 these principal analyses were: comparing C-B, comparing E-D and the critical analysis, comparing the levels of  
209 activation in the previous two comparisons; (C-B)-(E-D). The C-B comparison is a direct examination of our  
210 central prediction that activations in brain regions linked to prediction error would be greater for C presented  
211 alone, relative to B presented alone. The E-D comparison is similar to the C-B comparison but has a different

212 purpose. E and D serve as frequency matched controls to C and B, so any difference in the comparisons must be  
213 due to the presence or absence of the shared cue during training. The (C-B)-(E-D) comparison provides a direct  
214 test of these differences.

215 In addition to our principal analyses, we also conducted two further analyses. The first of these compared  
216 activation linked to AC and AB in brain areas previously linked to prediction error (and thus included in our  
217 ROI) in our training phase fMRI data. From both the behavioral data, and from the EXIT model, it is possible  
218 to predict that there will be more prediction errors on AC trials than AB trials, and hence areas associated  
219 with prediction error should be more active on AC trials than AB trials. The second of our additional analyses  
220 compared activation linked to BC and DE in our ROI during the test phase. EXIT does not predict a difference  
221 between these two compound cues; it instead predicts that the way attention is distributed between the cues  
222 within the compounds is the key difference. Nonetheless, as BC is the key behavioral cue, an obvious comparison  
223 to make is between BC and its frequency-matched control compound, DE. A further justification for this contrast  
224 is that theories other than EXIT might predict a neural difference between these two compounds.

225 The mask used for the ROI analysis was constructed using the WFU Pickatlas (Maldjian, Laurienti, Burdette,  
226 & Kraft, 2003), and was comprised of the brain regions we predicted to be linked to prediction error in our  
227 Introduction. Specifically, these regions were the bilateral caudate, putamen and nucleus accumbens, the right  
228 dorsolateral prefrontal cortex (BA 9 and BA 46), the medial anterior prefrontal cortex (BA 9 and BA 10) and the  
229 anterior cingulate (BA 24, BA 32 and BA 33). The number of voxels within this mask was 11952. Alongside  
230 ROI analysis, we also conducted exploratory whole brain analysis for each of the above comparisons.

231 The fMRI analyses were completed using a hierarchical general linear model, with first-level analyses  
232 conducted at the individual subject level and second-level analyses at the group level using a random effects  
233 model. The ROI analyses were conducted with a combined statistical threshold of  $p < .005$  and the following  
234 thresholds of contiguous voxels: 30 for the training phase analyses and 26 for the test phase analyses. These  
235 thresholds together produce an overall corrected threshold of  $p < .05$ ; based on cluster-level inference corrected  
236 for familywise error rate according to cluster size. These values were estimated using AlphaSim as implemented  
237 in the REST toolbox (Version 1.8, Song et al., 2011). For these calculations, smoothness was estimated within  
238 SPM12 using the group residuals from the general linear model and were 9.0 x 9.0 x 8.8 mm for the training  
239 phase and 9.7 x 9.7 x 9.4 mm for the test phase.

240 The test phase whole brain analyses were conducted with a combined statistical threshold of  $p < .001$   
241 and 110 contiguous voxels. These thresholds together produce an overall corrected threshold of  $p < .05$ ; again  
242 based on cluster-level inference corrected for familywise error rate according to cluster size. These values were  
243 again estimated using Alphasim (REST, Version 1.8, Song et al., 2011). For all analyses, normalized MNI

244 space coordinates were transformed to Talairach space using GingerALE (Eickhoff et al., 2011) and assigned  
245 anatomical labels using the Talairach Client (<http://talairach.org/client.html>) as per the atlas of Talairach and  
246 Tournoux (1988).

## 247 **3 Results**

### 248 **3.1 Behavioral Analyses**

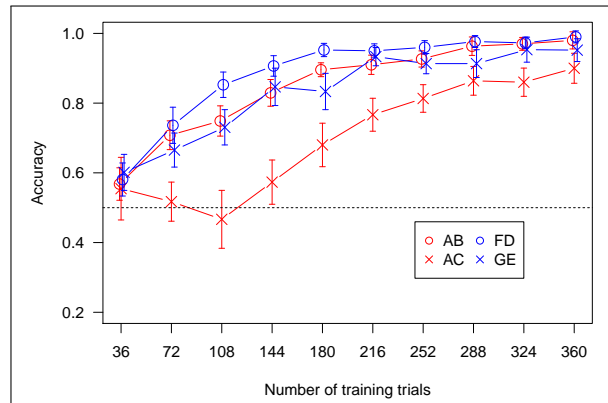
249 The accuracy of participants across the training phase is shown in Figure 2. A three-way ANOVA was  
250 conducted on the training phase data, looking at the effects of training block (first/last), stimulus frequency  
251 (common/rare) and shared cue (present/not present) on accuracy. Accuracy in the final block was significantly  
252 higher than the first block,  $F(1, 24) = 324.63, p < .001$ . No other significant main effects or interactions were  
253 found.

254 A further two-way ANOVA was conducted on the data in the final block of training, looking at the effects  
255 of stimulus frequency and shared cue on accuracy. Accuracy was significantly higher for the common stimulus  
256 compounds (AB and FD) than for the rare stimulus compounds (AC and GE),  $F(1, 24) = 5.23, p = .03$ . No  
257 other significant main effects or interactions were found.

258 Table 3 shows the response proportions for each of the stimuli presented in the test phase. The IBRE test  
259 stimulus BC was found to have a significantly greater proportion of rare responses than .5,  $BF_{10} = 31, t(24) =$   
260  $2.93, p = .003$ . Given there are only two response options in the current experiment, this demonstrates the  
261 presence of an IBRE. The proportion of common responses to the A stimulus was significantly greater than .5, as  
262 expected,  $t(24) = 6.14, p < .001$ . Also as expected, there were fewer rare responses to DE than to BC, although  
263 the evidence for this difference was inconclusive,  $BF_{10} = 1.8, t(24) = 1.57, p = .07$ .

264 Table 3 further shows the response proportions produced by the EXIT formal model (Kruschke, 2001b),  
265 within brackets next to the behavioral data. As can be seen from the Table, EXIT provides an extremely close fit  
266 to the behavioral data, capturing the response patterns for each stimulus,  $RMSD = .01, r^2 > .99$ . For technical

267



**Figure 2.** Training accuracy. The error bars are within-subject Cousineau-Morey 95% confidence intervals.

**Table 3.** Proportion of responses to each of the stimulus types presented in the test phase.

Stimulus type	Common	Rare
A	.76 (.76)	.24 (.24)
AB	.92 (.93)	.08 (.07)
AC	.19 (.17)	.81 (.83)
B	.92 (.90)	.08 (.10)
<b>BC</b>	.35 (.36)	<b>.65 (.64)</b>
C	.15 (.15)	.85 (.85)
D	.85 (.86)	.15 (.14)
<b>DE</b>	.44 (.43)	<b>.56 (.57)</b>
E	.24 (.24)	.76 (.76)
FD	.96 (.94)	.04 (.06)
GE	.11 (.13)	.89 (.87)

*Note.* Bold font indicates the behavioral results analysed. Values within brackets are simulated response proportions from the EXIT model.

268

## 270 3.2 Imaging Analyses

### 271 3.2.1 Training phase

272 We compared AC with AB in our ROI, during the training phase. This analysis (with thresholds of  $p < .005$   
 273 and 30 contiguous voxels) revealed a number of brain regions that exhibited greater activations for AC compared  
 274 to AB (see Figure 3). These regions were the bilateral caudate body (peak cluster size: 214, peak voxel  $x = -14$ ,  
 275  $y = 7$ ,  $z = 15$ ) and the right dorsolateral prefrontal cortex (BA 9; peak cluster size: 41, peak voxel  $x = 43$ ,  $y = 5$ ,  $z$   
 276  $= 32$ ).

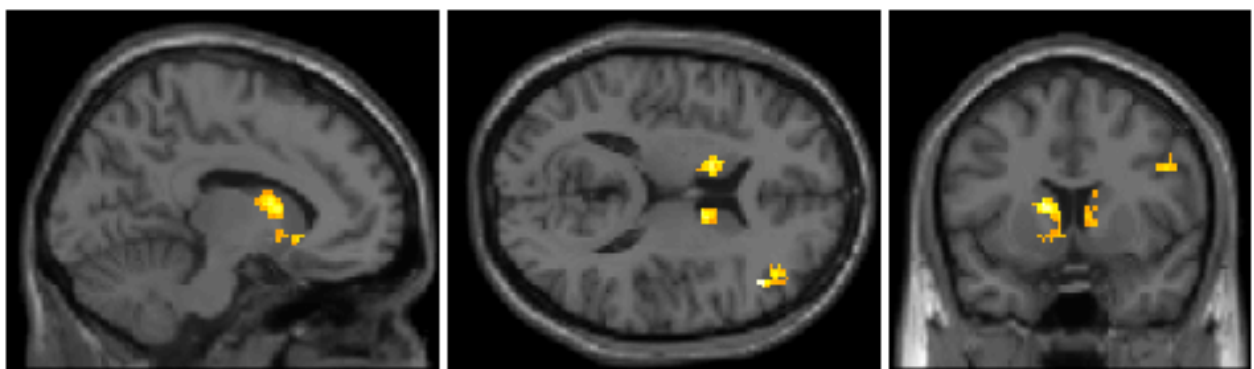
### 277 3.2.2 Test phase

278

279 **BC - DE comparison** The EXIT model does not predict a difference between these two compound cues,  
 280 because it is the distribution of attention within the compound that is predicted to vary between the two  
 281 compounds, not the total amount of attention to BC versus DE. Specifically, C is predicted to be more attended  
 282 than B, while attention should be more evenly distributed between D and E. As expected, no significant  
 283 differences were found, either in ROI or whole-brain analyses.

284 **C - B comparison** The ROI analysis (thresholds of  $p < .005$  and 26 contiguous voxels) revealed a number of  
 285 brain regions that exhibited greater activations for C (stimulus associated with the rare outcome) than for

286



$$x = -9, y = 10, z = 14$$

**Figure 3.** Areas that show greater activation for the AC cue compound compared to the AB cue compound under a ROI analysis, during the training phase. The thresholds used were  $p < .005$  and 30 contiguous voxels.

287 B (stimulus associated with the common outcome), see Figure 4 and Table 4. These regions included the

288 ventromedial prefrontal cortex (BA 10), medial prefrontal cortex (BA 9), right dorsolateral prefrontal cortex (BA  
289 9), bilateral caudate body, and left anterior cingulate (BA 32).

290 A number of brain areas included in the ROI analysis were also activated under whole brain analysis  
291 including a cluster comprising the right medial frontal cortex and the anterior cingulate (cluster size: 228, peak  
292 voxel  $x = 3, y = 55, z = 17$ ). Outside of brain areas already identified in the ROI analysis the right thalamus  
293 was activated (cluster size = 257, peak voxel  $x = 12, y = -11, z = 18$ ), as well as a separate cluster in the left  
294 cerebellum (cluster size = 111, peak voxel  $x = -25, y = -70, z = -28$ ).

295 **E - D comparison** The E-D comparison differs from the previous comparison in one key respect; the absence  
296 of a shared cue presented alongside E and D in training. Given the predictions of the error-driven learning  
297 account, and previous work (Kruschke, 2001a; Wills et al., 2014), we would expect to see no difference in  
298 activations here.

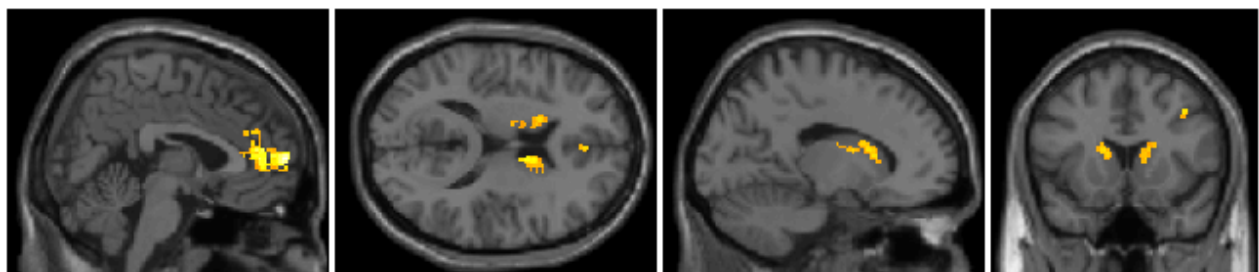
299 A ROI analysis examined activations for the E stimulus compared to the D stimulus and failed to find any  
300 areas that showed a significant difference in activation. Although this is unsurprising theoretically, these analyses  
301 were conducted to both stay consistent with the previous comparison and to characterize this comparison given  
302 its use in the final, critical, comparison. Whole brain analysis also failed to show any areas with a significant  
303 difference in activation.

304 **(C-B) - (E-D) comparison** This comparison is the critical analysis for the current experiment. The previous  
305 test phase comparisons differ in one key way; the presence or absence of a shared cue when training with those

**Table 4.** Brain regions activated during the test phase for a ROI analysis of C-B. The thresholds used were  $p < .005$  and 26 contiguous voxels.

Region	Cluster size	BA	Talairach coordinates			
			<i>x</i>	<i>y</i>	<i>z</i>	<i>z - score</i>
Right Ventromedial Prefrontal Cortex	219	10	3	55	17	4.46
Right Anterior Cingulate		32	3	39	15	3.88
Right Medial Prefrontal Cortex		9	3	48	18	3.58
Right Caudate Body	226		8	1	14	3.97
Right Caudate Body			12	-17	21	3.51
Right Caudate Body			14	2	23	3.12
Right Dorsolateral Prefrontal Cortex	32	6	34	6	41	3.43
Right Dorsolateral Prefrontal Cortex		9	39	10	38	2.82
Left Caudate Body	135		-8	1	10	3.26
Left Caudate Body			-16	8	17	3.24
Left Caudate Body			-12	14	13	3.08
Left Anterior Cingulate	58	32	-8	41	10	3.09
Left Ventromedial Prefrontal Cortex		10	-3	52	13	2.69

306



$x = 4, y = 13, z = 15$

$x = -13, y = 14, z = 5$

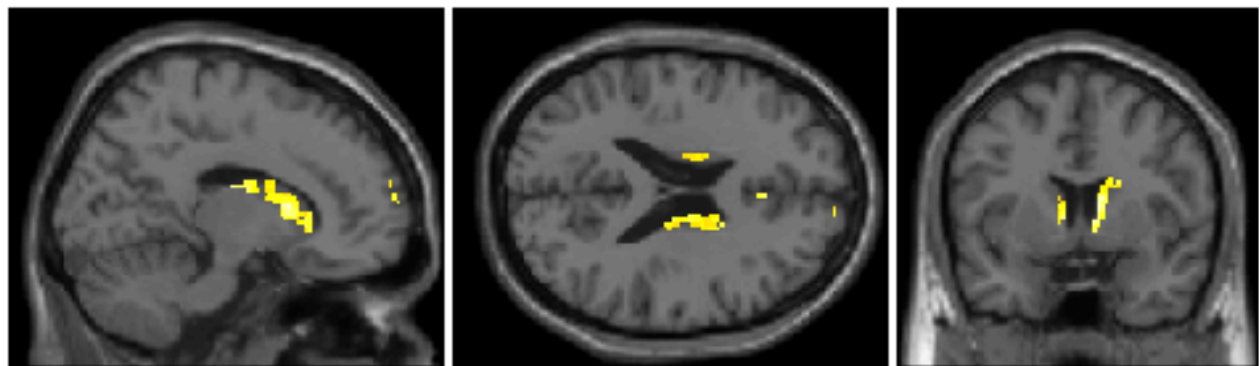
**Figure 4.** Areas that show greater activation for the C stimulus compared to the B stimulus during the test phase, under a ROI analysis. The thresholds used were  $p < .005$  and 26 contiguous voxels.

307

**Table 5.** Brain regions activated during the test phase for the ROI analysis of the comparison of the C-B comparison and the E-D comparison. The thresholds used were  $p < .005$  and 26 contiguous voxels.

Region	Cluster size	BA	Talairach coordinates			
			<i>x</i>	<i>y</i>	<i>z</i>	<i>z - score</i>
Right Caudate Body	395		8	1	14	3.86
Right Caudate Body			10	9	11	3.75
Right Caudate Body			6	6	5	3.55
Left Caudate Body	36		-8	1	10	3.45
Right Anterior Cingulate	32	24	4	21	24	3.39
Right Anterior Cingulate		24	4	29	18	3.06
Right Superior Prefrontal Cortex	45	9	8	56	24	3.35
Right Ventromedial Prefrontal Cortex		10	5	56	13	2.94
Left Anterior Cingulate	47	32	-8	39	11	3.17
Left Anterior Cingulate		32	-6	45	7	2.98
Left Caudate Body	32		-14	-11	19	3.13
Left Caudate Body	48		-16	8	17	3.05
Left Caudate Body			-16	16	15	2.75

308



309

$$x = 15, y = 7, z = 22$$

**Figure 5.** Areas that show greater activation for the C-B comparison compared to the E-D comparison under a ROI analysis, during the test phase. The thresholds used were  $p < .005$  and 26 contiguous voxels.



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

312 A ROI analysis revealed a number of brain regions exhibiting greater activation for the C-B comparison  
313 compared to the E-D comparison (Figure 5 and Table 5). Greater activation was noted in the bilateral caudate,  
314 the bilateral anterior cingulate, the right superior prefrontal cortex and right ventromedial prefrontal cortex.

315 The whole brain analysis also identified two clusters outside the areas identified in the ROI analysis, in the  
316 right thalamus (cluster size = 125, peak voxel  $x = 4, y = -19, z = 12$ ) and the left cerebellum (cluster size = 155,  
317 peak voxel  $x = -29, y = -68, z = -30$ ).

## 318 4 Discussion

319 The IBRE is a non-rational phenomenon in which people, having learned that cue compound AB predicts a  
320 common disease and cue compound AC predicts a rare disease, go on to predict that BC predicts the rare disease,  
321 in opposition to the underlying base rates (Kruschke, 1996; Medin & Edelson, 1988; Shanks, 1992). The current  
322 study was the first investigation of a successfully-observed IBRE with fMRI.

323 We made a number of predictions about brain activity and investigated them using ROI analysis. The  
324 predictions were made on the basis of: (1) an error-driven learning account of the IBRE, expressed as a formal  
325 model (Kruschke, 2001b), (2) a previous electrophysiological study of the IBRE (Wills et al., 2014), and (3) a  
326 substantial body of previous work on the neural correlates of prediction error (e.g. Fouragnan et al., 2018).

327 As predicted, a number of brain regions previously associated with prediction error during training showed  
328 greater activation during the test phase for the C cue relative to the B cue. These regions included the ventromedial  
329 prefrontal cortex, the medial prefrontal cortex, right dorsolateral prefrontal cortex, bilateral caudate body and  
330 left anterior cingulate. A number of previous studies have linked these areas to the occurrence of prediction error  
331 (e.g. Fletcher et al., 2001; Fouragnan et al., 2018; Garrison et al., 2013; Turner et al., 2004). These differences  
332 were not detectable for the frequency-matched control cues D and E, which were presented in training without  
333 the shared cue A. Greater activations were also noted in the right dorsolateral prefrontal cortex and bilateral  
334 caudate body during the training phase for the AC cue relative to the AB cue; a result consistent with both  
335 previous work and our test phase analysis.

336 Taken together, these results provide strong evidence in support of the prediction-error-based account of the  
337 IBRE (Kruschke, 2001b). Specifically, the current results, alongside those of Kruschke et al. (2005) and Wills et  
338 al. (2014), support the idea that the effects of prediction error during training persist into the test phase, and  
339 can be observed in singly-presented cues. These differences are characterized in EXIT as persistent changes in

340 attentional allocation, and this characterization in turn supports the idea that activity in brain areas associated  
341 with prediction error is sometimes associated with differences in attentional processing. Further support for  
342 Kruschke's account of the IBRE comes from the excellent level of quantitative fit of his EXIT model to the  
343 behavioral data of the present study (see Table 3).

344 Exploratory whole-brain analysis of the test phase identified several additional brain areas that might  
345 be involved in the IBRE. These areas were not predicted in advance so any inferences must be treated with  
346 some caution. One area in the thalamus showed a difference in activation for the C cue relative to the B cue.  
347 Given its role in relaying and processing sensory information (Schiff, 2008), its activation in this task is not  
348 unexpected. Another area in the left cerebellum also showed a difference in activation for the C cue relative to  
349 the B cue. This is perhaps unsurprising given that this area has been implicated in a wide range of cognitive  
350 tasks including learning (Desmond & Fiez, 1998); such as a previous category learning experiment (Carpenter,  
351 Wills, Benattayallah, & Milton, 2016).

352 There was some overlap between the areas of activation observed in the present work, and those observed  
353 in the only previous attempt to study the IBRE with fMRI (O'Bryan et al., 2018). O'Bryan et al. (2018) reported  
354 activations in the PFC, thalamus and cerebellum; areas also identified in our key contrast. Direct comparison of  
355 the two studies is difficult, however, due to differences in analysis methodology. The analyses conducted in the  
356 current study are direct stimulus contrasts, while O'Bryan et al. (2018) correlated brain activity with internal  
357 values of the dissimilarity-based extension of the Generalized Context Model (dissGCM; Stewart & Morin,  
358 2007). Nevertheless, the overlap in some of the regions identified across the studies is intriguing, even with this  
359 caveat in mind.

360 Inferring from this overlap should be approached with some caution though, as O'Bryan et al.'s conclusions  
361 appear somewhat different to those of the current study, and to those of a number of previous experiments on the  
362 IBRE. A key conclusion from O'Bryan et al. (2018)'s MVPA is that, on trials where participants respond rare  
363 to BC, they process B *more* intensively than C. O'Bryan et al. note that eye-tracking would be a good way to  
364 corroborate this finding; a methodology previously employed in the study of a variant of the IBRE by Kruschke  
365 et al. (2005). Kruschke et al. (2005) reported *less* attention to B than C on BC trials when an IBRE was observed,  
366 a finding further supported by the ERP results of Wills et al. (2014). Nonetheless, future work on the IBRE  
367 should further consider the theoretical implications of both sets of results..

368 In the current work, we have focussed on the predictions of the EXIT model, as these were the *a priori*  
369 basis of our experiment. Other formal models of category learning are available. One particularly pertinent  
370 alternative in the current case, given its predictions about the relationship between cognitive and neural processes,  
371 is the COmpetition between Verbal and Implicit Systems model (COVIS; Ashby, Alfonso-Reese, Turken, &

372 Waldron, 1998). We note that one of the areas identified in our key contrast was the caudate body, to which  
373 COVIS attributes stimulus representation in the procedural learning system. Nomura et al. (2007) suggest that  
374 feedback-driven learning strengthens synapses in the caudate through a reward signal, and the idea that the  
375 caudate is involved in some kind of associative learning process is consistent with a number of other related  
376 results (e.g. Carpenter et al., 2016; Seger & Cincotta, 2005). The COVIS procedural system, in its current form,  
377 does not provide an explanation for the IBRE, but it could potentially be modified to do so by the inclusion of  
378 the sort of error-driven attentional-allocation process employed in EXIT and investigated in the current work.

379 While we argue for the role of prediction error in the brain regions identified in our analysis, it is worth  
380 acknowledging that some of these areas, in particular the DLPFC, have been linked to other cognitive processes.  
381 Schlösser et al. (2009) evidenced a link between DLPFC and the processing of uncertainty; clearly this could play  
382 a role in the handling of the BC test cue, due to uncertainty generated as a result of the conflicting information  
383 provided by the B and C cues individually. Similarly, Badre and D'Esposito (2007, 2009) link the lateral PFC  
384 to hierarchical cognitive control processes, including attentional control. This is interesting, as EXIT arguably  
385 instantiates a controlled process of attentional reallocation; for example, it has previously been proposed that  
386 concurrent load disables attentional reallocation in this kind of model (Nosofsky & Kruschke, 2002).

## 387 **4.1 Conclusion**

388 The current study provides the first evidence linking the bilateral caudate body, the left anterior cingulate, the  
389 right dorsolateral prefrontal cortex, the ventromedial prefrontal cortex and medial prefrontal cortex to the IBRE.  
390 These neural correlates are strongly linked to the occurrence of prediction error; a concept implemented within  
391 the error-driven learning account of Kruschke (2001b). Therefore, this study both furthers the neuroscientific  
392 literature investigating prediction error and strongly supports the account implemented within Kruschke's (2001b)  
393 EXIT formal model.

## 394 **Notes**

395 Author contributions were as follows: ABI: Lead author on all aspects including write-up. FM: Assistance with  
396 fMRI analysis, interpretation, and write-up. CERE: Assistance with programming, data collection, and write-up  
397 AB: Radiography support. AJW: Experimental design, plus assistance with behavioral analysis, interpretation,  
398 and write-up.

## 399 **References**

400 Ashby, F., Alfonso-Reese, L., Turken, U., & Waldron, E. (1998). A neuropsychological theory of multiple

- 401 systems in category learning. *Psychological Review*, *105*, 442-481.
- 402 Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical  
403 organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *19*, 2082-2099.
- 404 Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews*  
405 *Neuroscience*, *10*, 659-669.
- 406 Baguley, T., & Kaye, D. (2010). Book review: Understanding psychology as a science: An introduction to  
407 scientific and statistical inference. *British Journal of Mathematical and Statistical Psychology*, *63*, 695-698.
- 408 Binder, A., & Estes, W. (1966). Transfer of response in visual recognition situations as a function of frequency  
409 variables. *Psychological Monographs: General and Applied*, *80*, 1-26.
- 410 Byrd, R., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization.  
411 *SIAM Journal on Scientific Computing*, *16*, 1190-1208.
- 412 Carpenter, K., Wills, A., Benattayallah, A., & Milton, F. (2016). A comparison of the neural correlates that  
413 underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*, 3557-3574.
- 414 Desmond, J., & Fiez, J. (1998). Neuroimaging studies of the cerebellum: language, learning and memory.  
415 *Trends in Cognitive Sciences*, *2*, 355-362.
- 416 Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological*  
417 *Science*, *6*, 274-290. doi: doi: 10.1177/1745691611406920
- 418 Eickhoff, S., Bzdok, D., Laird, A., Roski, C., S, C., Zilles, K., & Fox, P. (2011). Co-activation patterns  
419 distinguish cortical modules, their connectivity and functional differentiation. *Neuroimage*, *57*, 938-949.
- 420 FIL Methods Group, . (2014). SPM12 release notes [Computer software manual]. Retrieved from [https://](https://www.fil.ion.ucl.ac.uk/spm/software/spm12/)  
421 [www.fil.ion.ucl.ac.uk/spm/software/spm12/](https://www.fil.ion.ucl.ac.uk/spm/software/spm12/)
- 422 Fletcher, P., Anderson, J., Shanks, D., Honey, R., Carpenter, T., Donovan, T., ... Bullmore, E. (2001). Responses  
423 of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature*  
424 *Neuroscience*, *4*, 1043-1048.
- 425 Fouragnan, E., Retzler, C., & Philiastides, M. (2018). Separate neural representations of prediction error valence  
426 and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*, 2887-2906.
- 427 Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of  
428 neuroimaging studies. *Neuroscience and behavioural reviews*, *47*, 1297-1310.
- 429 Grinbrand, J., Erdeniz, T., Lindquist, M., Ferrera, V., & Hirsch, J. (2008). Detection of time-varying signals in  
430 event-related fMRI designs. *Neuroimage*, *43*, 509-520.
- 431 Haberg, G., Zito, G., Patria, F., & Sanes, J. (2001). Improved detection of event-related functional MRI signals  
432 using probability functions. *Neuroimage*, *14*, 1193-1205.
- 433 Inkster, A. B. (2019). *Attention, context and the inverse base rate effect*. (Doctoral dissertation, Plymouth

- 434 University, UK). Retrieved from <https://pearl.plymouth.ac.uk/handle/10026.1/14725>
- 435 Jeffreys, H. (1961). *The Theory of Probability* (3rd ed.). Oxford: Oxford University Press.
- 436 Juslin, P., Wennerholm, P., & Winman, A. (2001). High-level reasoning and base-rate use: Do we need  
437 cue-competition to explain the inverse base-rate effect? *Journal of Experimental Psychology: Learning,*  
438 *Memory, and Cognition, 27*, 849-871.
- 439 Kruschke, J. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory,*  
440 *and Cognition, 22*, 3-26.
- 441 Kruschke, J. (2001a). The inverse base-rate effect is not explained by eliminative inference. *Journal of*  
442 *Experimental Psychology: Learning, Memory, and Cognition, 27*, 1385-1400.
- 443 Kruschke, J. (2001b). Toward a unified model of attention in associative learning. *Journal of Mathematical*  
444 *Psychology, 45*, 812-863.
- 445 Kruschke, J. (2003). Attentional theory is a viable explanation of the inverse base rate effect: A reply to Winman,  
446 Wennerholm, and Juslin (2003). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*,  
447 1396-1400.
- 448 Kruschke, J., Kappenman, E., & Hetrick, W. (2005). Eye gaze and individual differences consistent with learned  
449 attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory,*  
450 *and Cognition, 31*, 830-845.
- 451 Mackintosh, N. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement.  
452 *Psychological Review, 82*, 417-421.
- 453 Maldjian, J., Laurienti, P., Burdette, J., & Kraft, R. (2003). An automated method for neuroanatomic and  
454 cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage, 19*, 1233-1239.
- 455 Medin, D., & Edelson, S. (1988). Problem structure and the use of base-rate information from experience.  
456 *Journal of Experimental Psychology: General, 117*, 68-85.
- 457 Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., . . . Reber, P. (2007). Neural correlates  
458 of rule-based and information-integration visual category learning. *Cerebral Cortex, 17*, 37-43.
- 459 Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning:  
460 Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review, 9*, 169-174.
- 461 O'Bryan, S., Worthy, D., Livesey, E., & Davis, T. (2018). Model-based fMRI reveals dissimilarity processes  
462 underlying base rate neglect. *eLife, 7*, e36395.
- 463 Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned  
464 but not of unconditioned stimuli. *Psychological review, 87*(6), 532.
- 465 R Core Team. (2018). R: A language and environment for statistical computing. [Computer software manual].  
466 Vienna, Austria. Retrieved from <https://www.R-project.org>

- 467 Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of  
468 reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning II: Current*  
469 *research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
- 470 Schiff, N. (2008). Central thalamic contributions to arousal regulation and neurological disorders of conscious-  
471 ness. *Annals of the New York Academy of Sciences*, *1129*, 105-118.
- 472 Schlösser, R., Nenadic, I., Wagner, G., Zysset, S., Koch, K., & Sauer, H. (2009). Dopaminergic modulation  
473 of brain systems subserving decision making under uncertainty: A study with fmri and methylphenidate  
474 challenge. *Synapse*, *63*, 429-442.
- 475 Seger, C., & Cincotta, C. (2005). The roles of the caudate nucleus in human classification learning. *Journal of*  
476 *Neuroscience*, *25*, 2941-2951.
- 477 Shanks, D. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*,  
478 *4*, 3-18.
- 479 Song, X., Dong, Z., Li, S., Zuo, X., Zhu, C., He, Y., ... Zang, Y. (2011). REST: A toolkit for resting-state  
480 functional magnetic resonance imaging data processing. *PLoS One*, *6*, e25031.
- 481 Stewart, N., & Morin, C. (2007). Dissimilarity is used as evidence of category membership in multidimensional  
482 perceptual categorization: a test of the similarity-dissimilarity generalized context model. *Quarterly Journal*  
483 *of Experimental Psychology*, *60*, 1337-1346.
- 484 Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of*  
485 *the ninth annual conference of the cognitive science society* (pp. 355–378).
- 486 Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional*  
487 *system: An approach to cerebral imaging*. Stuttgart: Thieme.
- 488 Turner, D., Aitken, M., Shanks, D., Sahakian, B., Scharzbauer, T. R. C., & Fletcher, P. (2004). The role of  
489 the lateral frontal cortex in causal associative learning: exploring preventative and super-learning. *Cerebral*  
490 *Cortex*, *14*, 872-880.
- 491 Wills, A., Dome, L., Edmunds, C., Honke, G., Inkster, A., Schlegelmilch, R., & Spicer, S. (2019). catlearn:  
492 Formal psychological models of categorization and learning. [Computer software manual]. Retrieved from  
493 <https://CRAN.R-project.org/package=catlearn> (R package version 0.6.2)
- 494 Wills, A., Lavric, A., Croft, G., & Hodgson, T. (2007). Predictive learning, prediction errors, and attention:  
495 Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience*, *19*, 843–854.
- 496 Wills, A., Lavric, A., Hemmings, Y., & Surrey, E. (2014). Attention, predictive learning, and the inverse  
497 base-rate effect: Evidence from event-related potentials. *NeuroImage*, *87*, 61-71.

## 498 **Appendix A: Modeling**

499 The simulation was conducted using *slpEXIT*, part of the *catlearn* R package (Wills et al., 2019). This  
500 implementation of EXIT is based on the model as described in Kruschke (2001b), with the inclusion of a bias cue  
501 that was later implemented in Kruschke (2003). The salience of the bias cue is represented by the  $\sigma$  parameter.

502 The EXIT model was applied to simulated training and test trials that replicated the details of the experi-  
503 mental procedure, generating response patterns for each simulated trial. The values of the free parameters given  
504 to the model were optimized using the *optim* function in R (R Core Team, 2018); specifically the limited memory  
505 Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Byrd, Lu, Nocedal, & Zhu, 1995). The sum of squared  
506 errors (SSE) between the model predictions and behavioral data was used as the objective function. As *optim*  
507 requires an initial set of starting parameters to vary, each free parameter within the EXIT model was initially set to  
508 one of two values. As there are 7 free parameters, this resulted in a total of  $2^7$  or 128 sets of parameter values. This  
509 produced 128 sets of optimized parameter values; the set with the lowest SSE was chosen. The parameter values  
510 within this final optimized set were:  $c = .746$ ,  $P = 2.383$ ,  $\phi = 2.963$ ,  $\lambda_g = .257$ ,  $\lambda_w = .047$ ,  $\lambda_x = 2.069$ ,  $\sigma = .031$ .