Learned predictiveness models predict opposite attention biases in the inverse base-rate effect

Hilary J. Don[1]

Tom Beesley[2]

Evan J. Livesey[1]

[1]The University of Sydney

[2]Lancaster University

Corresponding author:
Dr Hilary J. Don
hilary.don@tamu.edu
Department of Psychological and Brain Sciences,
Texas A&M University
4235 TAMU, College Station, TX 77840

Abstract

Several attention-based models of associative learning are built upon the learned predictiveness principle, whereby learning is optimised by attending to the most predictive features and ignoring the least predictive features. Despite their functional similarity, these models differ in their formal mechanisms, and thus may produce very different predictions in some circumstances. As we demonstrate, this is particularly evident in the inverse base-rate effect. Using simulations with a modified Mackintosh model and the EXIT model, we found that models based on the learned predictiveness principle can account for rare-outcome choice biases associated with the inverse base-rate effect, despite making *opposite* predictions for relative attention to rare versus common predictors. The models also make different predictions regarding changes in attention across training, and effects of context associations on attention to cues. Using a human causal learning task, we replicated the inverse base-rate effect and a recently reported reduction in this effect when the context is not predictive of the common outcome, and used eye-tracking to test model predictions about changes in attention both prior to making a decision, and during feedback. The results support the predictions made by EXIT, where the rare predictor commands greater attention than the common predictor throughout training. In addition, patterns of attention prior to making a decision differed to those during feedback, where effects of using a partially predictive context were evident only prior to making a prediction.

Keywords: attention; inverse base-rate effect; eye tracking; Mackintosh; EXIT; context learning

There is a popular assumption among theorists of associative learning that processes of attention in learning operate based on the *learned predictiveness* principle (e.g., Kruschke, 1996; 2001b; Mackintosh, 1975; Lovejoy, 1968; Sutherland & Mackintosh, 1971). This principle assumes that cues in our environment that reliably signal the occurrence of important or task-relevant outcomes increasingly capture attention as we learn about their predictive properties, and in turn become easier to learn about in the future because of the attention they command. Converging evidence from a range of psychological phenomena have confirmed this reciprocal relationship and highlighted its significance for cognitive psychology (see Le Pelley, Mitchell, Beesley, George, & Wills, 2016, for a review). This principle forms the basis for a class of attention-based models of associative learning. The most prominent of these models is the Mackintosh model (Mackintosh, 1975; hereafter simply referred to as *Mackintosh*).

Mackintosh offered a clear and persuasive explanation of the learned predictiveness principle as well as a formal description of the general operations by which learning processes may govern attention. These operations assume that attention increases to a cue that is a good predictor of the outcome relative to others presented at the same time, and attention decreases to poorer predictors. Both the general principle and specific operations outlined by Mackintosh have been critical in providing an explanation of related biases in learning, such as the learned predictiveness effect (e.g., Le Pelley & McLaren, 2003; Lochman & Wills, 2003), many of which are broadly consistent with the model. However, a recent review by Le Pelley et al. (2016) has brought into focus the fact that despite the considerable evidence consistent with the learned predictiveness principle, it is still not clear how learned changes in attention operate and, for instance, whether the processes

formalised by Mackintosh are indeed the most appropriate for capturing the effect. There is a need to assess the specific mechanisms offered by Mackintosh and other similar theories, especially in situations where they make diverging predictions. To this end, one phenomenon that may be particularly important for examining learned attention is the *inverse base-rate effect* (Medin & Edelson, 1988), a phenomenon that has conventionally been explained by learning models in which attentional biases are determined by associative prediction error, and yet has not been extensively used by associative learning theorists. Le Pelley et al. (2016) highlight the effect as potentially problematic for some attention models based on predictiveness principles, based on discrepancies in responses on single and compound cue trials. We will first introduce the inverse-base rate effect, and the predominant model of the effect, before returning to this issue in further detail.

The inverse base-rate effect

The *inverse base-rate effect* is a choice bias in human contingency learning. In the inverse base-rate effect, one frequently presented cue compound consistently predicts one outcome (e.g. AB-O1), while a less frequently presented cue compound predicts another outcome (e.g. AC-O2). Typically, AB-O1 and AC-O2 trials are presented in a 3:1 ratio such that cue B is relatively common compared to the rarer cue C and, likewise, the *base-rate* of O1 is high relative to O2. Cue A is therefore an *imperfect* predictor, as it is paired with both outcomes. Cue B (hereafter the *common predictor*) is a perfect predictor of the common outcome, O1 and cue C (hereafter the *rare predictor*) is a perfect predictor of the *rare* outcome, O2.[1] After learning these contingencies, participants are given a test phase in which they are presented with the

[1] Note that where there are multiple instantiations of the design, letters A–C will refer to all cues of the same type. That is, A refers to imperfect predictors, B to perfect predictors of common outcomes, and C to perfect predictors of rare outcomes.

individual cues, as well as several new combinations of cues, and asked to predict which outcome is most likely. When participants are shown the imperfect predictor (A) alone, participants tend to predict the common outcome, and to a lesser extent they show this same tendency on combined (ABC) trials. Although symptom A is associated with both outcomes, this response is consistent with the base-rates of the two outcomes. However when presented with the *conflicting* cue combination, BC, participants tend to predict the rare outcome, predicted by cue C. In this case, both cues are equally predictive of their respective outcomes, such that the specific cues do not provide evidence in favour of one outcome over the other. However, O1 occurs much more frequently than O2, and thus an arguably rational response, considering the differing base-rates, would be to predict O1 (Shanks, 1992). The inverse base-rate effect therefore refers to this seemingly irrational choice of the rare outcome on conflicting trials, and appears to be robust under a variety of task scenarios and stimuli (Dennis & Kruschke, 1998; Johansen, Fouquet & Shanks, 2010; Kalish, 2001; Kalish & Kruschke, 2000; Kruschke, Kappenman & Hetrick, 2005; Lamberts & Kent, 2007; Sherman et al., 2009).

The inverse base-rate effect is typically explained by prioritised attention to cue C during training (Kruschke, 1996; 2001a). This account suggests that AB-O1 trials are learned well and learned early, because they occur relatively frequently. As a result, both A and B become moderately associated with the common outcome, O1. The presence of A on the less frequent AC trials thus elicits an erroneous prediction of O1. To reduce subsequent error, and to preserve learning about AB trials, attention shifts away from the ambiguous cue A towards the more predictive cue, C. This attentional bias to C results in a stronger association between C and O2 than the association between B and O1. The change in attentional processing may also transfer

to BC trials, such that C is attended more than B and so cue C tends to control responding on BC trials at test. The idea that attention shifts away from A on AC trials is supported by the finding that a stronger association between A and the common outcome is necessary for rare outcome biases on BC trials (Shanks, 1992).

The attention account of the inverse base-rate effect has been formalised in the EXIT model (Kruschke, 2001a; 2001b), which is a connectionist model of error-driven attention and falls under the larger class of models based on the learned predictiveness principle. The EXIT model assumes that attention capacity is limited, and that attention is rapidly shifted towards cues that will reduce subsequent error. These attention distributions are then learned, and applied on subsequent trials according to their similarity to past exemplars. EXIT readily predicts the inverse base-rate effect and related *highlighting* effects (Kruschke, 1996; 2005), and as such, explanations of the effect have typically relied on the specific mechanisms proposed in the EXIT model. Kruschke (2001b) described the EXIT model as a connectionist implementation of the theoretical principles proposed by Mackintosh, and noted that the models make the same predictions for changes in attention. It is thus often assumed that the EXIT model operates similarly to Mackintosh, despite there being important differences (discussed below) in how attention changes in the two models. To date, this assumption remains largely untested. As such, the primary aim of this study was to formally compare the predictions that the models make about attention change using our recent data on the inverse base-rate effect as a test bed (Don & Livesey, 2017), and to test those predictions using measures of overt attention during learning and at test. We will briefly discuss three related issues that motivated choices about experimental design and simulations in the current study before discussing the models and simulations in further detail.

Discrepancies between single-cue and compound trials.

One previous study has raised questions about the ability of Mackintosh to account for some characteristics of the inverse base-rate effect. In an EEG study, Wills, Lavric, Hemmings & Surrey (2014) measured ERP correlates of selective attention when cues were presented individually at test. They found that, despite greater correlates of selective attention on C alone trials than B alone trials, common outcome responses to cue B alone were greater than rare outcome responses to cue C alone. To the best of our knowledge, this is the only time this difference in responding to B and C test trials has been tested statistically, yet this trend is also seen in several other cases (e.g. Bohil, Markman, & Maddox, 2005; Kruschke, 1996; Medin & Edelson, 1988; Medin & Bettger, 1991; Shanks, 1992; Winman, Wennerholm, Juslin, & Shanks, 2005; see Winman, Wennerholm & Juslin, 2003, for further discussion of this issue).

According to the Mackintosh model, this difference in accuracy for predictive cues implies greater associative strength for cue B than cue C, which should result in greater common outcome responses on conflicting BC trials. Yet, there were greater rare outcome responses on BC trials despite greater common responses on B trials than rare responses on C trials (Wills et al., 2014). This result is also problematic for a simple model of attention, in which the attention paid to a cue is directly related to its associative strength (Le Pelley et al., 2016). Nevertheless, Wills et al. (2014) suggest that some features of EXIT allow the model to account for the discrepancy between attention and responding to predictive cues. First, attention is normalised before influencing responding. That is, when cues are placed in direct competition (as in a compound trial), attention will influence the relative control of those cues over responding. However when a cue is presented individually, it has complete control

over responding, and attention will have little influence on responding. Thus responding on trials where B and C are presented individually may not be a good indication of the attention they receive when they are presented in compound. Attention in EXIT is also exemplar-mediated, such that the model can learn to direct attention away from A on AC trials, but maintain attention to A on AB trials. The similarity between AC and BC trials means that C should also receive prioritised attention on BC trials, which can account for an effect when associative strength for B may be higher than that for C (Kruschke, 2003; Wills et al., 2014).

The role of context learning

A potential way to reconcile the dissociation between attention and choice accuracy (e.g. Wills et al., 2014) with predictiveness principles is to assume a role of context learning. That is, the context may act as a cue that becomes associated with the outcomes, and subsequently influences responding on test trials. In an associative model, learning about the context is the primary mechanism for tracking overall base-rates irrespective of the predictive cues that are presented. As a result of the differences in base-rates for the two outcomes, the context will come to be more strongly associated with the common outcome. Therefore on B alone trials, both the cue and the context would predict the occurrence of the common outcome, whereas on C alone trials, context associations will act to weaken rare outcome predictions, even if C-O2 associations are stronger than B-O1 associations.[2] The effect of context

[2] Models of associative learning like Mackintosh often make the simple assumption that outcome predictions are based on a simple linear summation of the associative strengths of cues present on a given trial. Given this assumption, it may be difficult for these models to simultaneously predict that a) B + context could result in a stronger prediction of O1 than C + context of O2, and b) that B + C + context could lead to stronger prediction of O2 than O1. However, these circumstances could be possible if an assumption of nonlinearity in the summation process at test is made, for example, assuming that the context contributes more when only a single cue is present (e.g. B alone) than when two cues are present (e.g. BC). Mackintosh does not have a built-in capacity for this type of nonlinearity, which means that it may be limited in what it can predict, but also means its predictions are less parameter specific.

associations on attention to cues in the inverse base-rate effect has yet to be investigated.

In the EXIT model, context learning is captured by associations with a *bias node*, which may vary in salience. Initially EXIT was shown to predict higher accuracy for C alone trials than B alone trials in an overall fit of the data (Kruschke, 2001a). Yet, when EXIT was refit to the data with heavy weighting of the difference in B and C trial responses, it was able to predict a rare bias on BC trials when accuracy for B exceeded that for C (Kruschke, 2003). Notably, the salience of the bias node in this reweighted fit was high (.938) in comparison to the initial fit, which suggests that EXIT can account for the dissociation between responses on single and compound cue trials when the context is salient.[3] However, the inverse base-rate effect in this reweighted fit was reduced in magnitude compared to human choice, and the overall fit (indexed by root-mean-square error, RMSE) was poorer than in the initial, unweighted fit. We do not yet know whether the Mackintosh model can predict the inverse base-rate effect, or the dissociation in responding on single and compound cue trials.

Global outcome frequency effects

Assessing the influence of context associations in the inverse base-rate effect is difficult using the standard design alone, as the context will always be strongly associated with the frequently occurring outcome. However, we can assess the role of context by comparing the standard design to a "balanced" outcome design used in Don and Livesey (2017). This study compared the strength of the inverse base-rate effect with and without global outcome frequency differences. The design of the study

[3] The value of .938 is based on a fit to the data from Experiment 1 in Kruschke (1996). Kruschke (2003) states the bias salience in the initial, unweighted fit was .010, yet the bias salience reported in Kruschke (2001a) for the data from Experiment 1 of Kruschke (1996) is actually .401. A bias salience value of .00 is instead reported for a fit to the data from Experiment 2 in Kruschke (2001a).

is shown in Table 1. In the standard condition, O1 was always paired with the common compounds AB and DE, and O2 was always paired with the rare compounds AC and DF. Similarly, O3 was always paired with the common compounds GH and JK, while O4 was always paired with the rare compounds GI and JL. Thus, O1 and O3 were experienced three times as often as O2 and O4. In the balanced condition, each outcome was paired with both a common compound and a rare compound, such that all outcomes were experienced with equal frequency over the course of the experiment. For example, O1 was paired with the common compound AB and the rare compound DF, while O2 was paired with the rare AC and the common DE. Similarly, O3 was paired with the common compound GH and the rare compound JL, while O4 was paired with the rare GI and common JK. In this way, the *local* base-rate difference within each overlapping set is maintained (e.g. O1 was the common outcome within overlapping AB and AC trials, and O2 was the common outcome within DE and DF trials), but there is no *global* difference in the frequency of each outcome.

Table 1
Experimental Design Used in Experiment 2 of Don and Livesey (2017)

Phase	Group	Trial type	Base-rate	Trials			
Training	Standard	Common	3	AB – O1	DE – O1	GH – O3	JK – O3
		Rare	1	AC – O2	DF – O2	GI – O4	JL – O4
	Balanced	Common	3	AB – O1	DE – O2	GH – O3	JK – O4
		Rare	1	AC – O2	DF – O1	GI – O4	JL – O3
Test		Imperfect	1	A	D	G	J
		Conflicting	**1**	**BC**	**EF**	**HI**	**KL**
		Combined	1	ABC	DEF	GHI	JKL
		Common predictor	1	B	E	H	K
		Rare predictor	1	C	F	I	L
		Trained common	1	AB	DE	GH	JK
		Trained rare	1	AC	DF	GI	JL

Note: The critical conflicting test trials are indicated in bold.

Using this design, Don and Livesey (2017) found that the inverse base-rate effect (i.e. preference for the rare outcome on BC trials at test) was substantially reduced when each outcome had been experienced at an equal rate across the experiment. It is important, in the first instance, to demonstrate that attention-based associative learning models can actually account for this difference, as it may indicate that other non-associative decision processes play a critical role in producing the inverse base-rate effect. However, in principle, it should be possible for these models to do so by assuming that the associations between the context and the common outcome play a key role in enhancing the bias for the standard condition, where the outcome is globally common, but not in the balanced condition, where the outcome is only common for a given set of compounds. To test this idea, we first compare the predictions made by EXIT and Mackintosh in these designs, followed by an eye-tracking study to examine potential differences in attention to cues in the standard and balanced conditions.

Model simulations

To examine the predictions made by the EXIT and Mackintosh models, we fit both models to the choice data from the outcome frequency design of Experiment 2 in Don & Livesey (2017), and extracted the relevant attention weights across training.

EXIT and Mackintosh models

EXIT. The EXIT model used in this simulation is described in detail in Kruschke (2001a). In brief, when a stimulus is presented, corresponding cue nodes are activated. Cue nodes become associated with outcome nodes as a consequence of learning guided by prediction error, and cue node activation activates outcome nodes in order to generate predictions based on this learning. As part of this process, the cue

node activation also goes through a process of attentional reweighting. First, cue node activation spreads to exemplar nodes, which are activated to the extent that the presented cue combination is similar to the cue combination represented by the exemplar node. Exemplar node activation then spreads to attention gain nodes that competitively normalise attention. If the exemplar has been encountered previously, the attention gain nodes are activated based on prior learned attention distributions for that particular exemplar. Once attention gain is normalised, the resulting distribution of attention is combined multiplicatively with the original cue activations. This modified activation then spreads to the outcome nodes, where the model makes an outcome prediction based on the relative activation of outcome nodes. The model then provides corrective feedback. Attention is rapidly shifted towards cues that will reduce subsequent prediction error, and this change in attention is applied *before* the associations between cues and outcomes are updated. This is an important feature of the model, since the associative weight between a cue and an outcome is adjusted proportionally to the attention that is paid to that cue. Therefore, by adjusting attention prior to any changes in associative strength, there is an immediate effect of error-driven attention shifting on learning. Associative weights between exemplar and gain nodes are also updated, so that the new distribution of attention is used when the stimulus is encountered in the future. Full details of the free parameters used in the EXIT model can be found in the Appendix.

Mackintosh. The original version of the Mackintosh model—and the version that Wills et al. (2014) discuss critically in relation to the prediction it makes for single cues in the inverse base-rate effect—includes a separable error term for each cue. A separable error term limits the ability of the model to account for several learning phenomena, such as conditioned inhibition. Subsequent variations of

Mackintosh (e.g. Le Pelley, 2004; Suret & McLaren, 2005; Pearce & Mackintosh, 2010) instead use a summed error term, which can better capture cue competition or interactions between the predictions of cues (as in the case of conditioned inhibition). A version of Mackintosh with a summed error term may better account for the apparent discrepancy between attention and associative strength, and was therefore used in the following model fits.[4] Further details about this model can be found in the Appendix. In both the modified Mackintosh model and the EXIT model, attention to predictive cues should increase, and attention to non-predictive cues should decrease. Although both models operate on these similar theoretical principles, there are some important differences in the way each model operates. In Mackintosh, attention is synonymous with cue associability, such that the primary function of learned attention to a cue is to influence the rate of future learning about that cue. EXIT similarly assumes that more will be learned about attended cues. However in EXIT, attention influences output activation, such that cues with greater attention will have greater control over responding. Thus one clear difference is that in EXIT, the associative activation of the outputs can be influenced by the amount of attention cues receive. For instance, if cue A is associated with O1, the prediction of O1 can be enhanced when there is more attention to A and can be reduced when there is less attention to A. Mackintosh (1975) remained agnostic about the possibility of performance effects of this nature, but left these effects out of the model for the sake of simplicity. A second difference is that the learning of attention biases in EXIT is exemplar specific, such that attention to a particular cue may differ depending on the other cues with

[4] Using the same methods for assessing the EXIT and modified Mackintosh model in the current study, we found that the original Mackintosh model with a separable error term provided a poor fit of the choice data (RMSE = 13.50), and did not predict an inverse base-rate effect. Rather, it predicted a strong bias in choice for the common outcome on conflicting trials in the standard group, and no bias in the balanced group. Importantly, the original Mackintosh model made the same predictions for relative attention to cues as the modified Mackintosh model, described later in the paper.

which it is presented. For example, cue A may receive a different amount of attention when it is present on AB trials than when it is present on AC trials. In Mackintosh, alpha is stimulus-specific, rather than exemplar specific. Further, EXIT makes specific predictions about the timing of attention shifting. Specifically, attention shifts immediately to the most predictive cue, as a response to error, and associative weights are updated *after* this shift has occurred, while Mackintosh assumes that changes in attention are applied on subsequent trials.

Model fits were compared using two penalised-likelihood criteria; the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). While RMSE indicates the degree to which the model fits the data, AIC and BIC deal with the trade-off between goodness of fit and model complexity (the number of parameters) by penalising models for additional free parameters. For the purposes of model comparison, the model with the lowest AIC or BIC is regarded as the best fitting model, and the difference (ΔAIC, ΔBIC) indicates how well the best model performs in comparison to the other model. AIC and BIC were calculated from the residual sum squares (RSS) with the following equations, where k is the number of free parameters, and n is the number of data points:

$$\mathrm{AIC} = 2k + n\ln(RSS) \tag{6}$$

$$\mathrm{BIC} = n\ln(RSS/n) + k\ln(n) \tag{7}$$

Table 2
Outcome choice percentages from the test phase of Experiment 2 of Don & Livesey (2017), with predictions from EXIT and Mackintosh.

	Human				EXIT				Mackintosh			
	Standard		Balanced		Standard		Balanced		Standard		Balanced	
Trial	C	R	C	R	C	R	C	R	C	R	C	R
A	78.8	15.4	72.5	22.5	63.2	25.3	61.4	28.5	73.3	19.5	61.1	31.2
BC	**28.8**	**69.2**	**40.0**	**53.8**	**25.4**	**74.1**	**37.3**	**61.9**	**29.0**	**69.3**	**45.8**	**52.0**
ABC	42.3	57.1	64.4	33.8	40.2	59.4	57.5	42.2	43.5	56.3	63.1	36.7
Novel	65.4	26.9	58.1	30.0	62.1	25.3	60.4	28.4	73.3	19.5	61.1	31.2
B	**94.2**	**1.3**	**96.3**	**1.3**	**87.2**	**4.4**	**91.8**	**2.7**	**99.6**	**0.0**	**99.5**	**0.0**
C	**3.8**	**93.6**	**6.3**	**83.8**	**1.1**	**96.7**	**2.0**	**93.7**	**0.0**	**99.5**	**0.0**	**99.5**
AB	100	0.0	98.8	1.3	97.8	0.8	98.9	0.4	100	0.0	100	0.0
AC	5.1	92.9	8.1	90.0	0.4	99.1	1.1	97.6	0.0	99.9	0.0	99.9

Note: these data are presented in the percentage of outcome choice out of all possible outcome choices. C refers to common outcome responses, R refers to rare outcome responses. Trials of primary interest are presented in bold. Choice proportions do not necessarily sum to 100, as participants had a choice of four outcomes during the test phase, of which only two were the relevant common or relevant rare outcome.

Typically, ΔAIC of 0-2 indicates little difference between models, 4-7 indicates considerably less support for the model with larger AIC, and >10 indicates a great deal of support for the model with lower AIC (Burnham & Anderson, 2002).

Choice fits

Table 2 shows the choice data from Don & Livesey (2017, Experiment 2), and the predictions from EXIT and Mackintosh for each trial type. Both models provided a reasonable fit of the group differences, although overall, Mackintosh provided a considerably better fit of the data than EXIT (ΔAIC = 5.47; ΔBIC = 7.01). Both models were able to predict a reduction in the rare bias on conflicting trials in the balanced group compared to the standard group, as well as the accompanying decrease in the common bias on imperfect trials in the balanced group. However, the models differ in their predictions for individual predictive cue trials. The experimental data shows numerically greater accuracy for cue B alone than cue C alone. Although there were no significant group effects, this appears to be primarily driven by the balanced group. This might indicate that the reduced inverse base-rate effect in the

balanced group is due to better learning about the common predictor. Neither model predicts this result when using the parameters that provide the best overall fit. EXIT predicts greater rare choices for cue C alone than common choices for cue B alone in both groups, although this is somewhat weaker in the balanced group. In comparison, Mackintosh predicts that rare choices for C are equal to common choices for B in both groups.

Attention weights

After fitting the choice data, attention weights (α) for predictive cues, imperfect cues, and the context across training were extracted from the models. Figure 1 shows attention weights predicted by EXIT at two different stages, pre-shift attention (panel A and B; see equation 5 in Kruschke, 2001a) and post-shift attention (panel C and D), as well as the difference in post- and pre-shift attention (panel E and F).

In EXIT, post-shift attention indicates the amount of attention to cues after making an outcome prediction, at the point just prior to updating learning weights. These attention weights following feedback indicate which cues the model finds to be most useful in reducing error. It is worth noting that in this fit, the post-shift attention weights reach their asymptote very quickly, with little or no change after the second block of training. However, only some of this post-shift attention distribution is learned by the model and carried forward into the next trial, represented by pre-shift attention weights in panels A and B. This learned attention indicates the attention paid to cues *prior* to making a decision, and changes more gradually across training as it increments closer to the post-shift weights. This learned attention is represented by the pre-shift attention is panels A and B, which changes more gradually across training as it increments closer to the post-shift weights.

In both groups, post-shift attention (panel C and D) to predictive cues (B or C) is higher than the imperfect cue (A) on both common and rare trials. This difference in attention is much larger on rare trials than common trials, and this pattern did not differ substantially between the standard and balanced groups. Post-shift attention to the context did differ between groups, however. In the standard group, there was greater attention to the context on common trials than rare trials, consistent with the notion that the context was a more predictive cue on these trials. In contrast, the balanced group showed weaker post-shift attention to the context on common trials and stronger attention on rare trials.

Pre-shift attention (panel A and B) to predictive cues remained high throughout training in both groups, but there were differences in pre-shift attention to the imperfect cues between groups. In the standard group, there was a greater decrease in attention to the imperfect predictor on rare trials than common trials. In the balanced group, this difference in attention to the imperfect predictor on common and rare trials was comparatively reduced. Consequently, this might result in relatively stronger associative strength for the common predictor and weaker associative strength for the rare predictor in the balanced group compared to the standard group. Interestingly, the post-shift attention to the context does not appear to be learned, which may be due to the low bias salience in the best-fitting parameters.

Because not all post-shift attention is learned and carried forward to pre-shift attention on the next trial, we also plotted the difference between post-shift attention and pre-shift attention at each point in training (Panels E and F). This indicates the amount of change in attention to cues within a trial. Positive scores indicate a shift towards the cue, while negative scores indicate a shift away from the cue. The difference scores indicate that early in training, there are large shifts in attention away

from the imperfect predictor on rare trials, which then decrease across training as this attention distribution is learned.

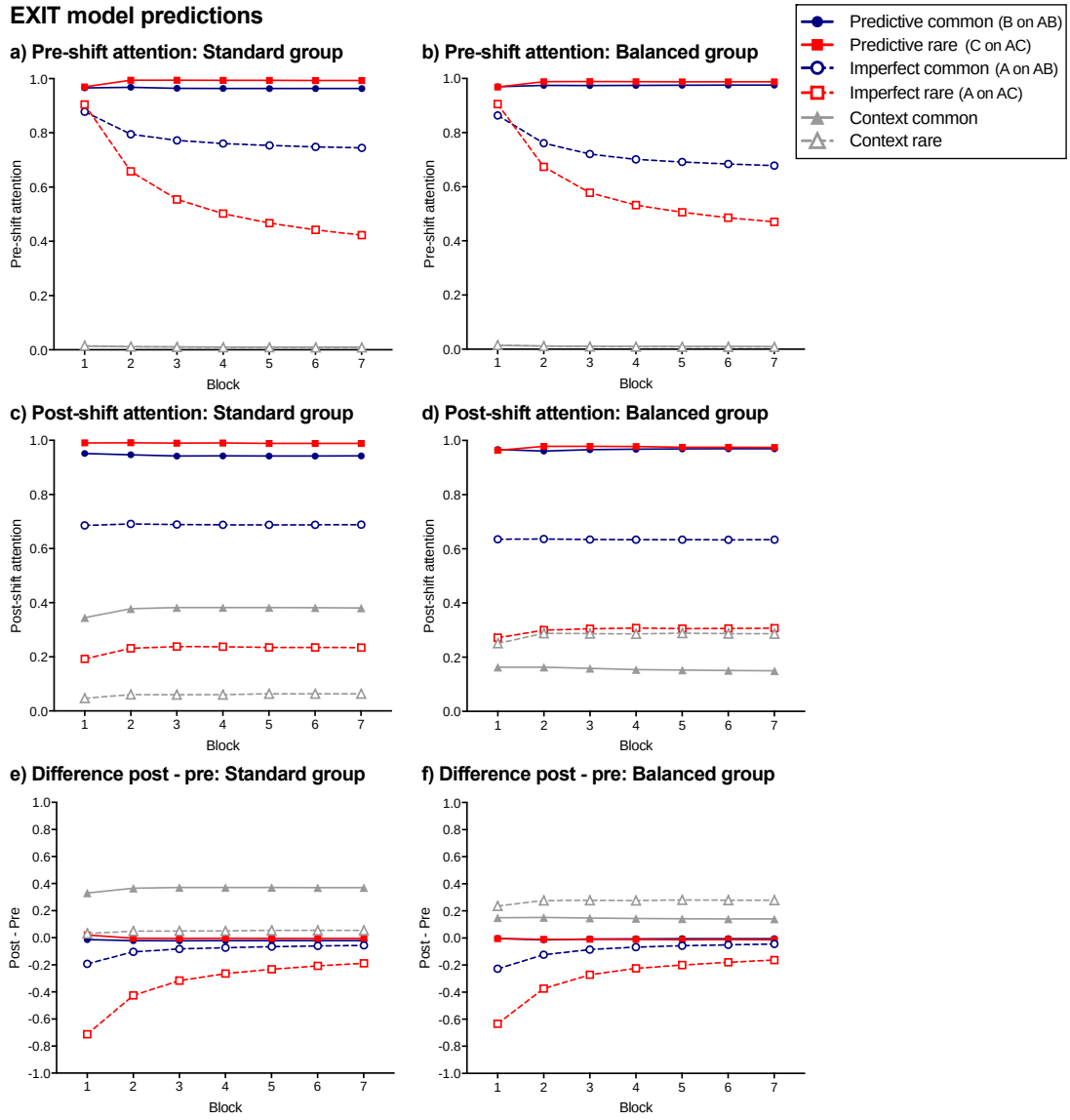

Figure 1. Pre-shift and post-shift attention for predictive cues, imperfect cues and context in Experiment 2 as predicted by EXIT.

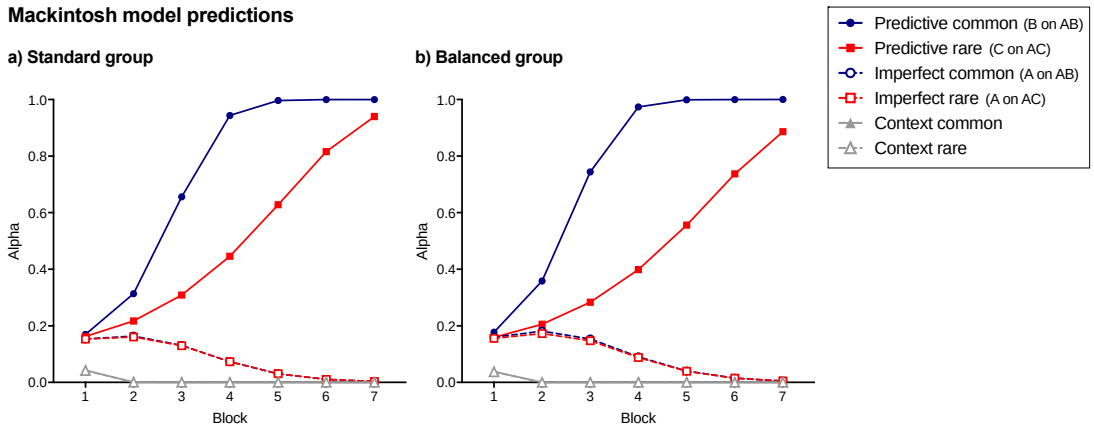

Figure 2. Alphas for predictive and imperfect cues in Experiment 2 as predicted by Mackintosh

Attention weights from the Mackintosh model are shown in Figure 2. Across training, α for predictive cues increased while α for the imperfect cues decreased. Overall, α for the predictive cue was higher on common trials than on rare trials. Due to the base-rates, B gains a higher α than the rare predictor early in training, while α for C gradually catches up across training. A modified Mackintosh model that contains a summed prediction error term can still predict an inverse base-rate effect on conflicting trials when α for B is greater than α for C, as there would be greater competition from A on common trials than rare trials. Although A is an imperfect predictor, it is a better predictor of the common outcome than the rare outcome, so it will compete effectively with B, but less so with C, such that C will gain a relatively large proportion of the associative strength with the rare outcome. In the balanced group, α is higher for B and lower for C compared to the standard group. Therefore Mackintosh appears to account for the difference between groups via greater overshadowing by A on rare trials, and weaker overshadowing by A on common trials.

Comparing predictions from EXIT and Mackintosh

Overall, the Mackintosh model provided a better fit of the choice data than EXIT, although the pattern of choice results was similar between models. Most interestingly, the predicted pattern of attention to cues during training varied substantially between the two models. The clearest and most critical difference is the predictions for relative attention paid to the predictive cues. The EXIT model predicted greater attention to cue C on rare AC trials than to cue B on common AB trials, whereas Mackintosh predicted greater attention to B on AB trials than to C on AC trials, at all points during training. These predictions are based on the optimal fit of parameters. However, it is important to consider whether these are general predictions from these models. This is particularly important because EXIT's operations are complex and nonlinear and other quite different combinations of parameters can satisfactorily predict an inverse base-rate effect (e.g. Kruschke, 2001a). We therefore ran 1000 simulations of each model with random parameters (within the constraints specified in the Appendix). For the EXIT model, 93.3% of simulations predicted a greater bias in attention to C on AC trials than B on AB trials in pre-shift attention, and 95.1% of parameters predicted this pattern in post-shift attention. For Mackintosh, 99.4% of random parameters predicted greater attention biases to B on AB trials than to C on AC trials. These differing ordinal predictions of attention to cues therefore appear to be parameter general. Mackintosh was also far less likely to predict an inverse base-rate effect on conflicting trials, producing the effect on only 9% of simulations, while EXIT produced the effect on 77% of simulations.[5]

[5] The original Mackintosh model with a separable error term predicted the same bias in relative attention as the modified Mackintosh model. Of 1000 simulations with random parameters, 86% predicted a greater bias to B on AB trials than to C on AC trials. None of these combinations of random parameters predicted an inverse base-rate effect.

The second difference is in changes in attention across training. Both models predict an increasing attention bias between predictive and non-predictive cues, however the influence of exemplar versus stimulus specific attention is clear. In EXIT, attention allocation to a specific cue is allowed to vary depending on the exemplar in which it is present. The trial base-rates affect attention biases because encountering a greater number of common AB trials results in a stronger prediction error on rare AC trials, thus leading to a stronger shift in attention towards the predictive cue on AC trials specifically. The attention bias developing across trials therefore manifests in EXIT as a compound-specific reduction in attention to A, which is stronger on AC trials than on AB trials. In Mackintosh, attention is stimulus specific, such that attention allocation to the imperfect cue on AC trials will be influenced by the attention it receives on AB trials, and vice versa. On both trial types, A is a poorer predictor, and therefore attention shifts towards the perfect predictor. Because there are three times as many AB trials than AC trials, this increase in attention to the perfect predictor will occur more frequently for AB trials than AC trials. The bias predicted by Mackintosh therefore manifests as a faster increase in attention for B than for C.

Finally, the two models also account for differences between standard and balanced groups in different ways. In EXIT, the bias for C over B was greater in the standard group than in the balanced group, while in Mackintosh, there was weaker attention to C in the balanced group than the standard group, although this difference is only subtle. EXIT also predicted differences in attention to the context between the standard and balanced groups. In the standard group, there was greater attention to the context on common trials than rare trials, which is a consequence of the context being a more useful predictor on common trials. This bias to context on common trials was

much weaker in the balanced group than the standard group, suggesting a less critical role of context when outcomes are experienced in equal frequency.

Testing model predictions with overt measures of attention

EXIT and Mackintosh make different ordinal predictions about relative attention to cues, how those attention biases change across training, and how attention biases differ based on global outcome frequency. We used eye-tracking to test these differing predictions. Eye gaze is often used as a measure of overt attention in learning (Beesley, Hanafi, Vadillo, Shanks, & Livesey, 2018; Beesley, Nguyen, Pearson & Le Pelley, 2015; Easdale, Le Pelley & Beesley, 2018; Le Pelley, Beesley & Griffiths, 2014; Rehder & Hoffman, 2005; Thorwart, Livesey, Wilhelm, Liu & Lachnit, 2017, Wills, Lavric, Croft & Hodgson, 2007). While it is possible to make covert shifts of attention without accompanying eye movements, attention and gaze are generally closely related (Posner, 1980). Although neither model specifies the way in which attention will translate to overall eye gaze, the models at least make different ordinal predictions in terms of which cues should experience greater attention. We assume that measures of fixation time will provide an indication of how attention is allocated to cues and how that allocation changes throughout training.

To the best of our knowledge, this is the first study reporting measures of overt attention in the inverse base-rate effect, although researchers have used eye-tracking to study the related highlighting effect, where AB trials are trained prior to the introduction of AC trials (Kruschke, Kappenman & Hetrick, 2005). They found greater fixation time for cue C on BC test trials, and a greater bias in fixation time to the predictive cue on AC trials than AB trials at test. It might be tempting to assume that this provides clear evidence in favour of EXIT's predictions for attention. Yet,

differences in attention to cues across training were not reported, and again, it is unknown whether base-rate differences when all trials are experienced at each stage of training will result in the same differences in attention as when cues are trained in a staged manner as in highlighting. It is also unclear whether choice on BC trials is a result of prioritised attention to C, leading to greater control by cue C, or whether it is due to greater learning about the C-O2 association during training. This could be important for isolating the locus of the inverse base-rate effect and the cause of accuracy differences for individual (B and C) cue trials.

There are currently inconsistent results regarding which cue is processed to a greater extent on BC trials at test. Wills et al. (2014) found greater ERP correlates of selective attention for C than for B when they were presented individually at test. In a recent fMRI study O'Bryan, Worthy, Livesey, & Davis (2018) used multivoxel pattern analysis to determine the extent to which participants were activating information involved in the representation of common and rare predictors. Faces, objects and scenes were used as cues, as these categories have well-defined regions of representation in the cortex. For example, imperfect predictors were always faces, but common and rare predictors were either objects or scenes, balanced within subjects. Prior to training, regions of sensitivity to these categories were determined for each participant, which were then compared to patterns of activation on conflicting test trials. This technique provides an index of neural similarity, which reflects the strength of representation of specific features, and is assumed to index a combination of feature-based attention and memory retrieval for what has been learned about those features. O'Bryan et al. found neural activity indicative of greater representation of cue B than cue C on conflicting trials when participants chose the rare outcome. These neural measures of stimulus processing are quite different from one another in

their time course and founding assumptions, and also distinctly different from measures of overt attention typically used in learning. Nevertheless they are certainly sufficient to warrant examining fixation time to cues at test, broken down according to outcome choice.

The EXIT model in particular makes clear predictions about attention shifts after an outcome choice, when the correct outcome is revealed. That is, attention is driven towards the cue that is most likely to reduce future error before the end of the trial. Part of this attention shift is learned, such that a proportion of the new attention distribution is applied when a similar stimulus is encountered. Due to this distinction, measures of fixation time were divided into two time periods; fixation time from the onset of cues until a response is made, which should reflect the learned pre-shift attention to cues in Figure 1 (panels A and B), and fixation time after making a response, during corrective feedback, which should reflect post-shift attention illustrated in Figure 1. It is possible that overt gaze at this stage of the trial will reflect the end-state of attention (as in panels C and D), or the amount of updating of attention biases required between pre-shift attention and end-state attention, indicated by the change in attention from pre-shift to post-shift attention (as in panels E & F).

In this experiment, participants completed either the standard or balanced version of base-rate training, followed by the typical inverse base-rate effect transfer test trials (see Table 3). Note that in this experiment, we used only two outcomes, rather than the four outcomes that were used in Don & Livesey (2017). The contingencies were presented using an allergist task, which has been used frequently in human contingency learning studies (e.g., Larkin, Aitken & Dickinson, 1998; Le Pelley & McLaren, 2001; Van Hamme & Wasserman, 1994; Waldmann & Holyoak, 1992) and has proven useful for studying attention transfer in the learned

predictiveness effect (Don & Livesey, 2015; Le Pelley & McLaren, 2003; Shone, Harris & Livesey, 2015). EXIT predicts greater pre-decision attention to cue C on AC trials than to cue B on AB trials. This bias should be greater in the standard group than in the balanced group, and should increase over training. During feedback, there should also be a stronger bias in gaze towards the predictive cue on rare trials than common trials, which does not differ between groups. If attention during feedback reflects end-state attention, this gaze bias should be acquired early and remain consistent throughout training, but if it reflects the change from pre- to post-shift attention, the gaze bias should decrease across training. Mackintosh instead predicts greater attention to cue B on AB trials than to cue C on AC trials, and that this difference between the two predictors decreases across training. Mackintosh does not make specific predictions about attention biases during the feedback period of the trial, though one might assume that post-feedback eye gaze could reflect the updating of attention that occurs post-learning. If this were the case then one would expect stronger attention biases towards B early in training, and stronger attention towards C later in training, as alpha for B reaches ceiling faster than alpha for C, and prediction error on AB trials reaches floor much faster than on AC trials.

Method

Participants

Ethical approval for this study was obtained from the Human Research Ethics Committee at the University of Sydney. Sixty-four undergraduate students from the University of Sydney participated in return for partial course credit (51 female, mean age = 19.6 years, *SD* = 4.21). Participants were randomly allocated to standard and balanced conditions (*n* = 32). Because there was no precedent for the expected effect

size in attention, we chose a sample size that was greater than that used in other studies examining eye-gaze in learning (e.g., Easdale, Le Pelley & Beesley, 2018; Le Pelley, Beesley & Griffiths, 2011). We continued collecting data until we reached 32 participants per group.

Apparatus and stimuli

The experiment was programmed using PsychToolbox for Matlab (Kleiner, Brainard & Pelli, 2007). Participants were tested individually, and eye gaze was measured using a 23-inch Tobii TX300 Eye Tracker, which has a sample rate of 300 Hz. Participants were seated approximately 55cm from the monitor, with a chin rest to maintain a fixed position. The eye tracker was calibrated using a five-point procedure at the beginning of the experiment. Cue stimuli were 300 x 300 pixel images of *Coffee*, *Fish*, *Lemon*, *Cheese*, *Eggs*, *Garlic*, *Bread*, *Peanuts*, *Avocado*, *Banana*, *Bacon*, *Peas*, *Apple*, *Mushrooms*, *Strawberries*, *Broccoli*, *Cherries*, *Butter*, *Olive Oil*, *Chocolate*, *Carrots*, *Peach*, *Milk*, and *Prawns*, randomly allocated to cues A-L. These were presented horizontally aligned on the upper half of the screen. The two outcome stimuli were randomly allocated from the allergic reactions *Headache*, *Nausea*, *Rash* and *Fever*, presented in text boxes on the lower half of the screen, vertically aligned with the center of the two cues. Stimuli were presented following a 500ms presentation of a fixation cross at the horizontal centre of the screen. The distance between the centre of the cross and the centre of each cue was 15cm. Outcome options appeared 500ms after the presentation of the cues. Responses were made using a standard mouse and keyboard. Feedback was provided for two seconds in the centre of the screen.

Table 3
Experimental Design for the Current Study

Phase	Group	Trial type	Base-rate	Trials			
Training	Standard	Common	3	AB – O1	DE – O1	GH – O1	JK – O1
		Rare	1	AC – O2	DF – O2	GI – O2	JL – O2
	Balanced	Common	3	AB – O1	DE – O2	GH – O1	JK – O2
		Rare	1	AC – O2	DF – O1	GI – O2	JL – O1
Test		Imperfect	1	A	D	G	J
		Conflicting	**1**	**BC**	**EF**	**HI**	**KL**
		Combined	1	ABC	DEF	GHI	JKL
		Common predictor	1	B	E	H	K
		Rare predictor	1	C	F	I	L
		Trained common	1	AB	DE	GH	JK
		Trained rare	1	AC	DF	GI	JL

Note: The critical conflicting test trials are indicated in bold.

Procedure

Participants assumed the role of a doctor whose task was to determine which foods were causing which allergic reaction in their patient, Mr X. On each trial during training, two food cues appeared, and participants were asked to predict which allergic reaction would occur after Mr X had eaten that meal by clicking on one of two outcome options on the lower half of the screen. Once an outcome was selected, the options disappeared and corrective feedback was provided. Participants were told that at first they would have to guess, but using the feedback provided, their accuracy should improve over time. The arrangement of the two cues on the screen was counterbalanced within each block and the positions of the two outcomes were randomised for each participant but held constant throughout the session. There were seven blocks of training including the contingencies presented in Table 3, in a 3:1 base-rate. Each block contained six presentations of the frequent trial types and two presentations of the infrequent trial types, such that there were 224 training trials.

The test phase began immediately following training. Participants were instructed to use the knowledge that they had gained so far to predict the allergic

reaction that was most likely to occur after Mr X ate meals containing one, two, or three foods. In this phase, food cues appeared on the upper half of the screen and participants made their prediction by clicking on an outcome, which then turned blue. Once a choice was made, a continuous rating scale ranging from "not at all confident" to "very confident" appeared beneath the options, and participants were asked to rate their confidence that they had made the correct choice. Responding was self-paced, and participants were able to modify their response before moving to the next trial. The test phase included one repetition of each of the transfer trials in Table 3, presented in random order. The arrangement of the two cues on the screen was randomised for each trial, and outcome options were presented in the same arrangement as in training.

Eye gaze analysis

Fixation time on each trial was separated into two time periods – a "pre-decision" period which spanned stimulus onset to a response, and a "feedback" period, which began once a prediction had been made, and continued while the feedback was presented on screen until the end of the trial. On each trial, the percentage of missing samples was calculated, and the data from the eye with the least missing samples were used. Gaps of missing data less than 75 milliseconds were interpolated between the data preceding and following the gap. Fixations were determined by a displacement method (Salvucci & Goldberg, 2000). The horizontal and vertical coordinates of gaze data were analysed in 150ms windows, and a fixation was determined if the coordinates did not deviate beyond a range of 75 pixels. This window was then extended until a deviation of greater than 75 pixels was recorded, to determine fixation length. Fixation position was taken as the mean horizontal and vertical pixel coordinates across the fixation sample. Any fixation recorded within a

500 x 500 pixel region of interest (ROI) centered on the cue image (providing a 100 pixel ROI surrounding the cue image) was taken as a fixation on that cue..

Trials without any recorded fixations were removed from the analyses, and participants with more than 30% missing trials in either the decision or feedback period in the training, or during the test phase, were excluded from the respective eye gaze analysis (their data were used in the behavioural analyses). This resulted in the exclusion of one participant from the training phase gaze analysis, and two participants from the test phase analysis. For the remaining participants, mean excluded trials were 3.7%, (*SE* = 0.8) during training, and 0.8% (*SE* = 0.4) during the test phase. To control for potential differences in response time between trial types, fixation time was analysed as a proportion of response time in the pre-decision period in training, and as a proportion of the time taken to first select an outcome during the test phase. Fixations while rating confidence, or altering outcome choice during the test phase were not included in the analysis.

Results

Choice responses

Training accuracy. To analyse the training accuracy data shown in Figure 3A, a 2 x (2) x (7) mixed measures ANOVA was run with group (standard vs. balanced) as a between subjects factor, and trial type (common vs. rare) and block (1 – 7) as within subjects factors. There was a significant linear effect of block, $F(1,62) = 303.76$, $p < .001$, $\eta_p^2 = .83$, and quadratic effect of block, $F(1,62) = 166.82$, $p < .001$, $\eta_p^2 = .729$. There was a significant main effect of trial type, with greater accuracy for common trials overall, $F(1,62) = 134.57$, $p < .001$, $\eta_p^2 = .685$, and group, with greater accuracy in the standard group overall, $F(1,62) = 7.33$, $p = .009$, $\eta_p^2 = .106$. There was

a significant interaction between group and trial type, such that there was a greater difference in accuracy for common and rare trials for the balanced group, $F(1,62) = 10.59$, $p = .002$, $\eta_p^2 = .146$. There were also significant interactions between the linear effect of block and trial type, $F(1,62) = 73.29$, $p < .001$, $\eta_p^2 = .542$, and between the quadratic effect of block and trial type, $F(1,62) = 24.74$, $p < .001$, $\eta_p^2 = .285$, indicating common trials were learned about faster than rare trials. There were also interactions between the linear effect of block and group, $F(1,62) = 10.25$, $p = .002$, $\eta_p^2 = .142$, and a significant three-way interaction between the linear effect of block, group, and trial type, where there was a greater difference in the speed of learning for common and rare trials for the balanced group than the standard group, $F(1,62) = 7.07$, $p = .01$, $\eta_p^2 = .102$.

Test. Responses for all trial types are shown in Table 4, and the proportion of relevant rare outcome choices for the three critical transfer trials is shown in Figure 3B. Although the combined trials are included here for consistency with previous research, we do not place too much weight on them, as response biases on combined trials are generally much less reliable than those seen on imperfect and conflicting trials (see Shanks, 1992). Where a null effect is of potential theoretical importance, we include Bayes Factor (BF) to assess the evidence in favour of the null, based on Rouder, Speckman, Sun, Morey & Iverson's (2009) JZS prior with scaling factor r = .707.

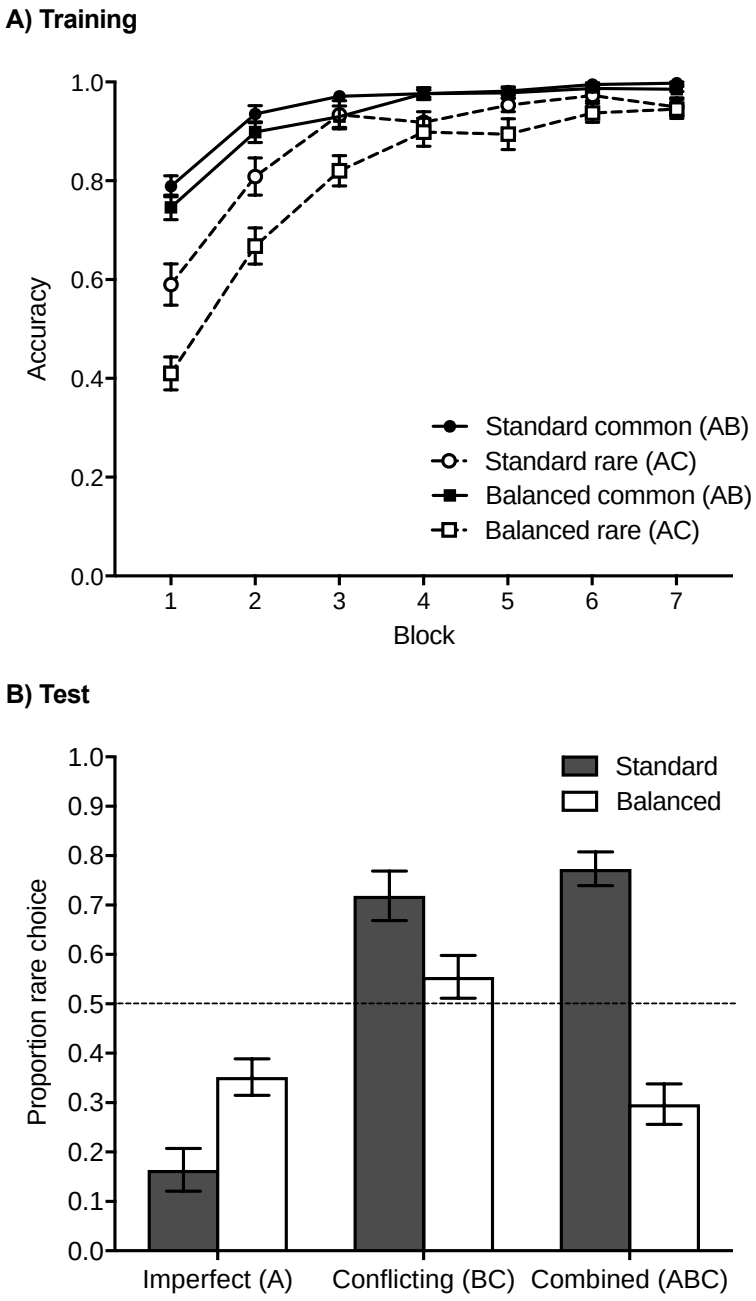

Figure 3. Choice data including A) Response accuracy during training for common and rare trial types in each group, and B) Proportion rare choice on imperfect, conflicting and combined transfer trials during the test phase. Error bars indicate standard error of the mean.

Table 4.
Choice responses and confidence ratings for all test trials

		Choice		
Transfer Trial	Group	Common	Rare	Confidence
Imperfect	Standard	.836	.164	52.89
A, D, G, J	Balanced	.648	.352	57.41
Conflicting	Standard	.281	.719	60.20
BC, EF, HI, KL	Balanced	.445	.555	51.75
Combined	Standard	.227	.773	68.48
ABC, DEF, GHI, JKL	Balanced	.703	.297	65.12
Common predictor	Standard	.953	.047	76.76
B, E, H, K	Balanced	.961	.039	79.22
Rare predictor	Standard	.039	.961	78.36
C, F, I, L	Balanced	.094	.906	75.23
Trained Common	Standard	.984	.016	94.61
AB, DE, GH, JK	Balanced	.984	.016	94.87
Trained Rare	Standard	.023	.977	94.28
AC, DF, GI, JL	Balanced	.070	.930	88.03

Note: as there are only two outcomes in this experiment, choice proportions sum to 1.

The proportion of rare choice on these trials replicates the pattern of results reported in Don and Livesey (2017; Experiment 2) fairly closely. There was a significant inverse base-rate effect in the standard group with a greater proportion of rare outcome choices on conflicting (BC) trials, $t(31) = 4.39$, $p < .001$, $d = 0.78$. While there was a small numerical bias for choosing the rare outcome on conflicting trials in the balanced group, this did not reach significance, $t(31) = 1.27$, $p = .214$, $d = 0.22$ The rare-bias was significantly weaker on conflicting trials in the balanced than the standard group, $t(62) = 2.49$, $p = .016$, $d = 0.62$. On imperfect (A) trials, responding was significantly common-biased in both groups, lowest $t(31) = 4.01$, $p < .001$, $d = 0.71$, but this was significantly weaker in the balanced than standard group, $t(62) = 3.31$, $p = .002$, $d = 0.83$. On combined (ABC) trials, choice was significantly rare-

biased in the standard group, $t(31) = 7.96$, $p < .001$, $d = 1.41$, and significantly common-biased in the balanced group, $t(31) = 4.94$, $p < .001$, $d = 0.87$, and there was a significant difference in choices between the two groups, $t(62) = 8.89$, $p < .001$, $d = 2.22$. Overall, there was no significant difference in accuracy for common (mean = .96, *SD* = .12) and rare predictors (mean correct responses = .93, *SD* = .16), $F(1,62) = 1.32$, $p = .255$, $\eta_p^2 = .021$, $BF_{01} = 3.95$. There was also no main effect or interaction with group, highest $F(1,62) = 2.35$, $p = .131$, $\eta_p^2 = .036$.

Eye gaze

Pre-decision. Pre-decision fixation time on each cue, as a proportion of total decision time is shown in Figure 4. We ran a 2 x (2) x (2) x (7) repeated measures ANOVA with group as a between-subjects factor, and cue predictiveness (imperfect vs. perfect), trial type (common vs. rare) and block (1-7) as within-subjects factors. There were significantly longer fixations on predictive cues than imperfect cues overall, $F(1,61) = 21.97$, $p < .001$, $\eta_p^2 = .265$. Fixation time was also generally higher on rare trials than on common trials, $F(1,61) = 12.18$, $p = .001$, $\eta_p^2 = .166$, but a significant interaction with linear trend in block revealed that this difference decreased over training, $F(1,61) = 14.94$, $p = .001$, $\eta_p^2 = .197$. Critically, there was a significant predictiveness x trial type interaction $F(1,61) = 12.50$, $p = .001$, $\eta_p^2 = .17$, with a greater bias for perfect predictors on rare trials than on common trials. This result is clearly consistent with the predictions of EXIT and not Mackintosh. There was also a significant three-way interaction between predictiveness, trial type and group, $F(1,61) = 8.02$, $p = .006$, $\eta_p^2 = .12$. Further analysis for each group separately showed that in the standard group, there was a significant interaction between predictiveness and trial type, $F(1,30) = 22.09$, $p < .001$, $\eta_p^2 = .424$, where there was a significant bias towards the predictive cue on rare trials, $F(1,30) = 21.93$, $p < .001$, η_p^2

= .422, but not on common trials, $F < 1$, $BF_{01} = 5.21$. In contrast, in the balanced group, the bias towards the predictive cue did not differ between common and rare trials, $F < 1$, $BF_{01} = 4.67$, and the predictive cue bias was significant for both trial types, lowest $F(1,30) = 5.88$, $p = .021$, $\eta_p^2 = .159$. The bias for rare predictors on rare trials did not differ between standard and balanced groups $F(1,59) = 1.71$, $p = .196$, $\eta_p^2 = .027$, $BF_{01} = 1.90$.

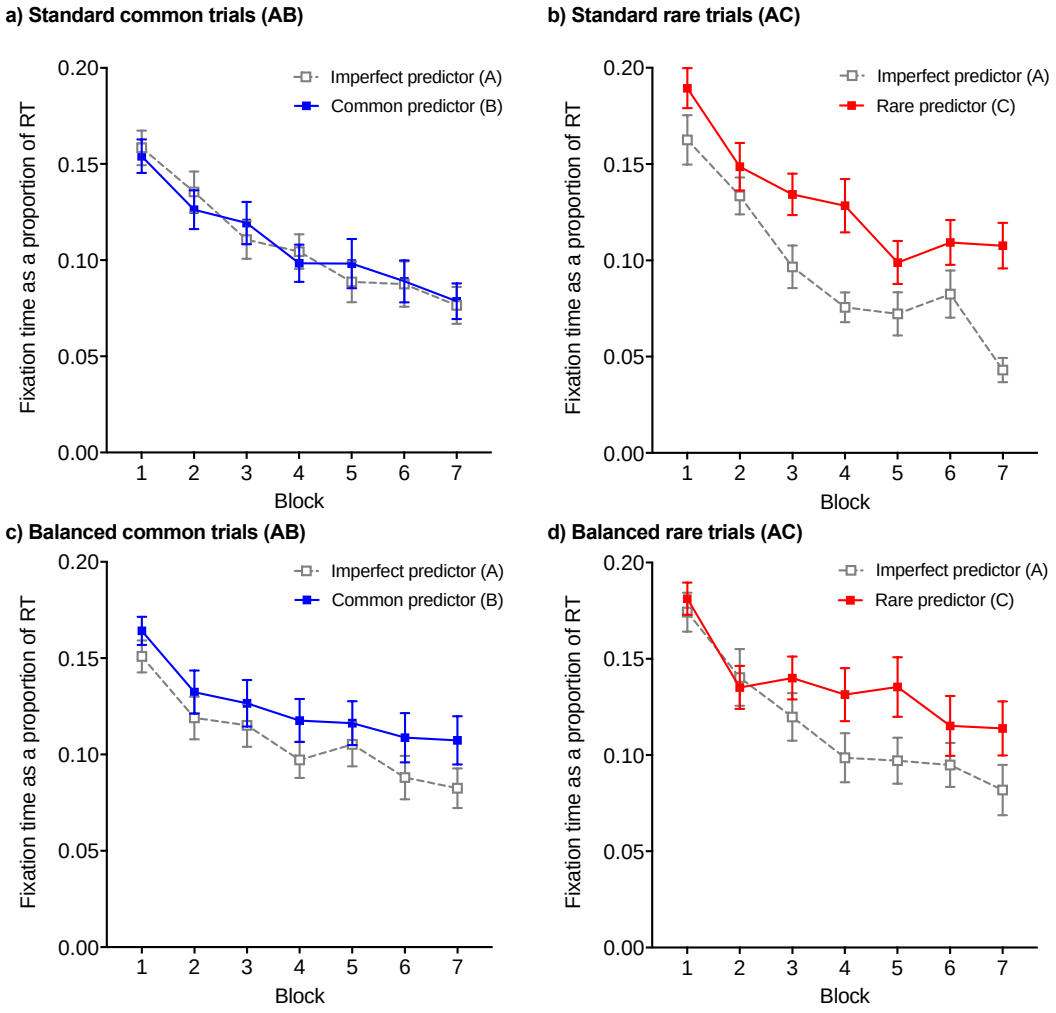

Figure 4. Fixation time as a proportion of response time (RT) during the pre-decision period for A) common trials in the standard group, B) rare trials in the standard group, C) common trials in the balanced group, and D) rare trials in the balanced group. Error bars indicate standard error of the mean.

To determine whether biases in attention changed over the course of training, we examined linear effects of block. This revealed a significant interaction between block, predictiveness and trial type, $F(1,61) = 5.98$, $p = .017$, $\eta_p^2 = .089$. To further investigate this interaction, common and rare trials were analysed separately. The bias for predictive cues relative to imperfect predictors significantly increased across training on rare trials, $F(1,61) = 11.20$, $p = .001$, $\eta_p^2 = .155$, but there was no significant change on common trials, $F(1,61) = 3.49$, $p = .067$, $\eta_p^2 = .054$. These effects did not interact with group, Fs < 1.

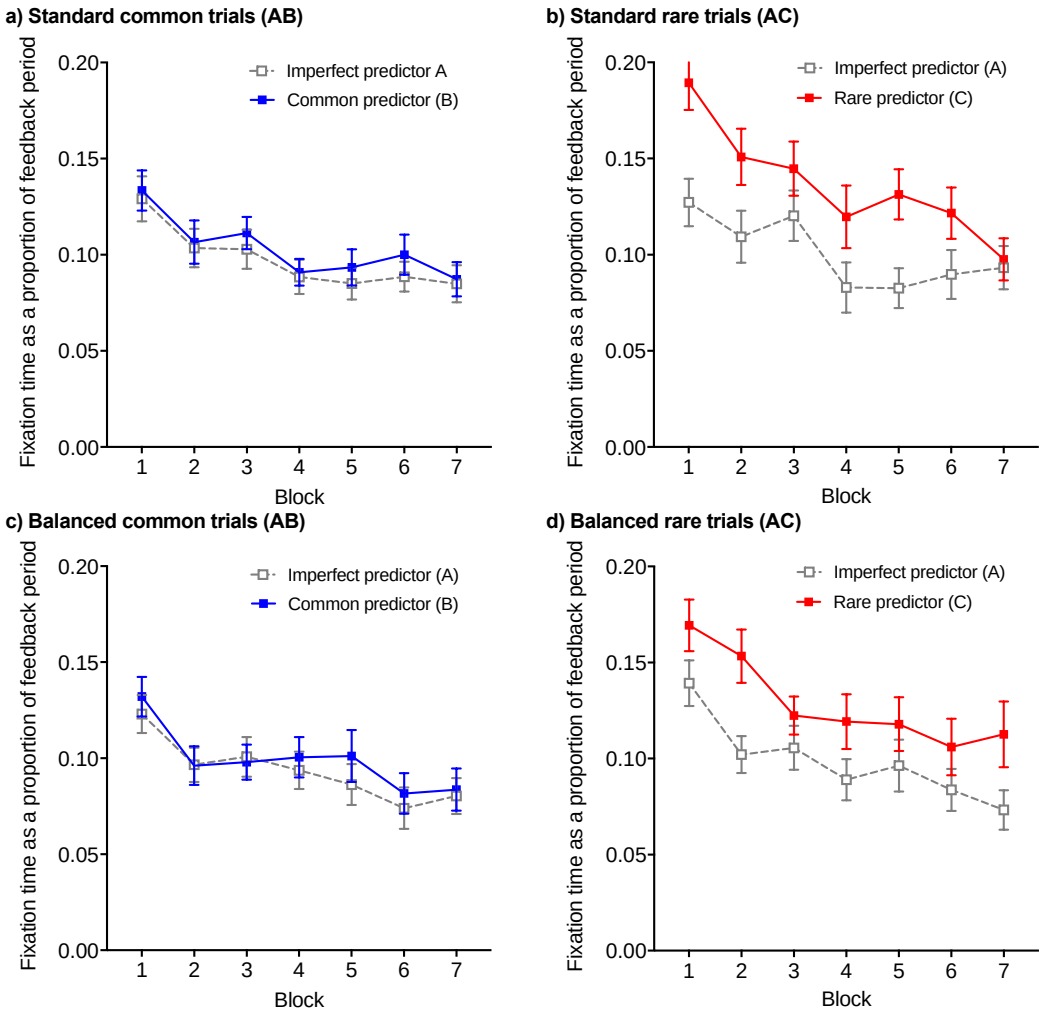

Figure 5. Fixation time as a proportion of feedback time for a) common trials in the standard group, b) rare trials in the standard group, c) common trials in the balanced group, and d) rare trials in the balanced group. Error bars indicate standard error of the mean

Feedback. Fixation time during feedback, shown in Figure 5, was analysed as a proportion of the fixed feedback time of two seconds. There were again significantly longer fixations on predictive cues than imperfect cues, $F(1,61) = 28.32$, $p < .001$, $\eta_p^2 = .317$. There was also greater fixation time on rare trials than common trials overall, $F(1,61) = 44.65$, $p < .001$, $\eta_p^2 = .423$, which decreased over training, $F(1,61) = 8.15$, $p = .006$, $\eta_p^2 = .118$. There was a significant interaction between trial type and cue type, $F(1,61) = 20.30$, $p < .001$, $\eta_p^2 = .250$, such that there was a greater difference in fixation time to predictive and imperfect cues on rare trials than on common trials. This effect of cue predictiveness was only significant on rare trials, $F(1,61) = 31.02$, $p < .001$, $\eta_p^2 = .337$, and not on common trials, $F(1,61) = 3.29$, $p = .075$, $\eta_p^2 = .051$, $BF_{01} = 1.52$.

Unlike fixations during the decision period, the interaction between cue predictiveness and trial type did not further interact with group, $F < 1$. There was a significant three-way interaction between the linear trend of block, predictiveness and trial type, $F(1,61) = 4.56$, $p = .037$, $\eta_p^2 = .07$. This indicates that the bias towards the predictive cue decreased across training for rare trials, $F(1,61) = 4.22$, $p = .044$, $\eta_p^2 = .065$, but not for common trials, $F < 1$. There was no significant interaction with group for either trial type, highest $F(1,61) = 2.31$, $p = .134$, $\eta_p^2 = .037$.

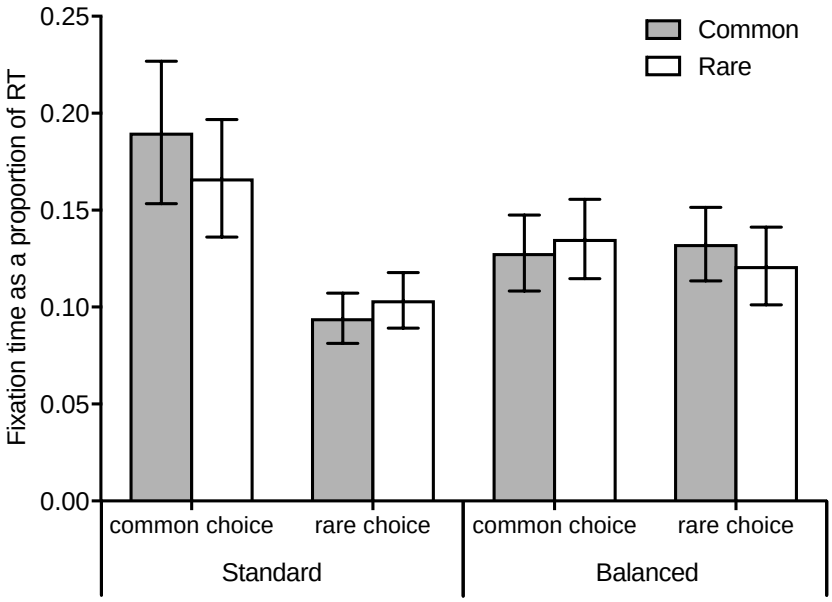

Figure 6. Gaze to common and rare predictors according to outcome choice on conflicting trials during the test phase.

Test. On conflicting test trials, there was no difference in fixation time to common or rare predictors, $F < 1$, $BF_{01} = 4.74$, and no effects of group, $Fs < 1$, lowest $BF_{01} = 3.50$. There was also no difference in attention to common and rare cues when considering only the trials on which a rare outcome was chosen, $F < 1$, $BF_{01} = 7.09$ (see Figure 6; cf. O'Bryan et al., 2017), and no main effect or interaction with group, highest $F(1,43) = 1.94$, $p = .171$, $\eta_p^2 = 0.043$, $BF_{01} = 1.72$.

Discussion

This study tested the conflicting predictions of EXIT and Mackintosh about attention to cues in the inverse base-rate effect, and the influence of context associations on attention to cues. Overall, the results provided greater relative evidence for the predictions of EXIT over Mackintosh. The experiment also replicated an effect of global outcome frequency on the inverse base-rate effect. The bias towards choosing the rare outcome on conflicting trials was significantly weaker

in the balanced group where the outcome occurred at an equal rate across the entire experiment, suggesting an influence of context associations on choice biases.

Overall, there was a greater fixation bias to cue C on rare trials than to cue B on common trials, and this effect differed between the two training procedures. In the standard group, both cues were attended equally on AB trials, but there was greater attention to C on AC trials. In the balanced group, more time was spent attending to predictive cues B or C than to A regardless of the trial type. During the test phase, there were no differences in fixation time to common and rare cues on conflicting trials, even when only considering trials where the rare outcome was chosen. This pattern also did not differ between groups.

Previous studies have shown that choice accuracy for B is greater than choice accuracy for C at test, in the presence of an inverse base-rate effect (e.g. Wills et al., 2014). In this experiment, there was no benefit in accuracy for B over C when tested individually, and this did not differ between standard and balanced groups. However, responding on these trials was close to ceiling and therefore a reliable difference in performance on these trials may be difficult to detect. The following discussion will compare the results with model predictions.

Relative attention to cues

The EXIT model predicted a greater bias in attention to C on AC trials than to B on AB trials, whereas Mackintosh predicted the opposite pattern, with higher attention to B on AB trials than to C on AC trials. Measures of fixation time indicated a greater bias for C over A than for B over A, both prior to making an outcome choice (in the standard group only), and during feedback (in both groups). These results indicate a clear attention advantage for C over B, and are more consistent with the predictions from EXIT.

Changes in attention over training

There were differences in the change in attention biases throughout training between the decision period and the feedback period. During the decision period, eye-gaze biases towards C on rare AC trials became more pronounced over training, and there was no significant change in biases on common trials. This pattern of attention is mostly consistent with the pattern of pre-shift attention biases predicted by EXIT, and likely reflects the current state of learning about cues. That is, attention prior to choice is an exploitative process, in which attention is directed towards cues that will be most useful in producing a correct outcome prediction.

During feedback, preferential attention to C on AC trials was stronger earlier in training than later in training. This change in attention appears to follow the reduction in prediction error, which was higher early in training, and lower later in training. Thus, attention during feedback appears to be a response to error, directed towards the cues most likely to reduce future prediction error.

Interestingly, the reduction in this bias for the rare predictor during feedback is also broadly consistent with a metric we derived from EXIT; the amount of attention change from pre-shift attention to post-shift attention, which also declines over training (Figure 1, Panels E and F). Thus attention during feedback may reflect an updating process, where attention biases are corrected between what was attended prior to making a choice, and an optimal distribution of attention. For example, if C is not well attended during the decision period, then a greater amount of attention may be allocated to that cue in feedback, compared to when C is well attended during the decision period.

Currently, gaze fixation during feedback is not widely used or reported in studies of attention in learning. Future research should further test whether attention

during this time period is indeed reflective of prediction error. If this is the case, gaze fixation during feedback will provide a useful measure of updating processes in future work.

The role of context learning

The comparison of standard and balanced conditions in these experiments allows us to determine the influence of context associations on the inverse base-rate effect, and on attention to predictive cues. The difference in frequency of overlapping compounds in both groups means that the imperfect predictor (e.g. A in AB and AC) should be associated with the relevant common outcome within each compound pair. However, in the standard condition, one outcome is consistently paired with common compounds, such that the context is also more strongly associated with that outcome. In the balanced condition, each outcome is paired with both common and rare compounds, such that the context would be equally associated with both outcomes. Consistent with this idea, the EXIT model predicted greater attention to the context on common trials than on rare trials in the standard group, but not in the balanced group, which suggests a greater influence of context learning in the standard group. The Mackintosh model instead predicted that the context loses attention rapidly in both groups.

The influence of context associations can provide a reasonable explanation for the differences in choice and pre-decision attention biases between groups. In the standard condition, participants can rely on the base-rate to help make accurate predictions on common AB trials. Although B is more predictive than A, it may be less necessary to focus on either cue in particular, relative to the balanced group in which the overall base-rate is not helpful. In the balanced group, the context does not provide a good prediction of either outcome, and therefore more attention would need

to be paid to the most predictive cues on every trial (cue B on common trials, and cue C on rare trials), in order to make an accurate response. Consistent with this idea, the common-bias to A was weaker in the balanced group, suggesting that they were paying less attention to this cue than in the standard group. Several authors have noted an association between the strength of common responses to imperfect cues and the strength of the inverse base-rate effect (e.g. Shanks, 1992; Kruschke et al., 2005), which has been taken as support for attention accounts. That is, the more A is associated with O1, the more attention should be shifted away from A to C on AC trials. Although we observed an equal pre-decision gaze bias to rare predictors on AC trials in both groups, increased attention to B on AB trials in the balanced group would result in a weaker association between A and the common outcome.[6] As such, the reduced rare outcome bias on conflicting trials in the balanced group may well be a result of differences in attention due to context associations. That is, an equivalent predictive cue bias on AB and AC trials might result in cue C having less relative control over responding compared to cue B on BC trials.

Again consistent with the predictions of EXIT, the pattern of post-feedback attention was no different between groups. Choice differences between groups therefore appear to be more strongly reflected in the pre-decision attention to cues, rather than error-driven attention shifts during feedback.

Attention at test

Our fixation time data from the test phase does not support either of the previous results regarding the strength of cue processing at test (Kruschke et al., 2005; O'Bryan et al., 2018). The absence of fixation time differences suggests that the effect

[6] It is worth noting however that the inverse base-rate effect likely cannot be explained by appealing to context learning alone. In Experiment 3 of Don & Livesey (2017), AB-O1 and AC-O2 were trained in equal frequency, while pairings of high-frequency filler trials with O1 provided a strong overall base-rate difference, and therefore strong context-O1 associations. These conditions were insufficient to produce the inverse base-rate effect on BC trials.

is not driven by *continued* overt attention biases to the rare predictor into the test phase. Rather, the contribution of attention to choice appears to be constrained to initial learning of the contingencies. Although researchers have emphasised the importance of attention biases at test for the inverse base-rate effect (e.g. Wills et al., 2014; O'Bryan et al., 2018), the EXIT model does not necessarily require continued attention biases at test to predict an inverse base-rate effect. Indeed the predecessor to EXIT, the ADIT model (Kruschke, 1996), was able to predict the inverse base-rate effect in the absence of learned attention to cues simply by relying on rapid attention shifts driven by prediction error on a given trial but not retained across trials.

Rule based-processes in the inverse base-rate effect

The purpose of this paper was to explore how well attention-based learning models can account for the inverse base-rate effect, and indeed it is clear that we have found changes in attention that are likely to contribute to choice biases. While we have focused primarily on attentional explanations of the inverse base-rate effect, the effect is not universally explained in these terms. There is a broad literature on the role of rules and reasoning in human learning (e.g., Mitchell, De Houwer & Lovibond, 2009), and the inverse base-rate effect specifically (Juslin, Wennerholm & Winman, 2001; Winman, Wennerholm, Juslin & Shanks, 2005). As we have previously discussed (Don & Livesey, 2017), non-attentional inferential processes may play a role in the choice of the rare outcome on conflicting trials, as well as differences in choice between standard and balanced groups. The most comprehensive inferential explanation for the inverse base-rate effect describes the rare bias as a consequence of eliminating the common outcome because of the noticeable dissimilarity between AB-O1 training trials and BC test trials (Juslin et al., 2001). There is now evidence from a range of studies that this eliminative inferential account

fails to capture many of the properties of the effect and makes corollary predictions that are demonstrably incorrect (see Don & Livesey, 2017 for a discussion). Although inferential hypotheses cannot provide a full account of the effect, we have found here that attention models are also limited in accounting for all eye-gaze and choice effects. There is good reason to assume that learned attentional biases are not fully controlled by deliberate inferential cognitive processes (e.g. Cobos, Vadillo, Luque & Le Pelley, 2018; Don & Livesey, 2015; Shone, Harris & Livesey, 2015), and applying associative learning algorithms to model attentional change remains a viable and justifiable approach, especially given the wealth of new empirical evidence supporting a link between associative learning and biases in visual attention (e.g., Failing & Theeuwes, 2018; Feldmann-Wüstefeld, Uengoer & Schubö, 2015; Livesey, Harris, & Harris, 2009; Luque, Vadillo, Gutiérrez-Cobo & Le Pelley, 2018). Nevertheless, by the same token, it seems likely that participants' choices at test are a result of a combination of processes including not only associative memory retrieval, but also inferential reasoning of some kind. There is certainly evidence of this in similar learning tasks that are designed to tease apart such influences when participants are faced with a novel test trial that requires generalization from trained instances (e.g. Don, Goldwater, Otto, & Livesey, 2016; Shanks & Darby, 1998; Wills, Graham, Koh, McLaren & Rolland, 2011). This may be the reason why both learning models do a less than impressive job of capturing the pattern of choice data in its entirety.

Associability versus attention

Attention is a notoriously vague term in psychology, and while it is consistently used in learning theoretic circles to refer to stimulus selectivity, there are many forms of stimulus selection to which it can be applied. The processes governing

each selection mechanism may not be the same. Classic attentional models of conditioning such as Mackintosh (1975) sought to characterise changes in stimulus *associability* specifically, that is the rate at which a cue enters into new learning, determined by how processing of that cue is prioritised over others. Mackintosh was explicitly agnostic about the possibility that these changes may also affect performance, for instance enhancing or diminishing the impact of previously learned associations on behaviour or indeed manifesting in overt attentive behaviours such as gaze fixation.

This is a key difference between Mackintosh's theory and Kruschke's EXIT model, in which attention explicitly affects both associability *and* performance, and it gives us cause to be cautious about ruling out the Mackintosh model's account entirely. Although several studies have found that predictive cues possess higher associability *and* attract longer dwell times than non-predictive cues (e.g. Le Pelley et al., 2011; Thorwart et al., 2017), beyond this correlation between two biases, it is not necessarily clear how associability is related to eye-gaze. While our eye-gaze results are clearly inconsistent with predictions derived from Mackintosh, it is possible that cue associability follows a different pattern, possibly one that is more consistent with Mackintosh than EXIT. It will be important for future research to determine whether cue associability changes in the same way as the eye-gaze measures reported here. For instance, in the tradition of the learned predictiveness effect, future research should test whether differences in the base-rates of the cue-outcome pairings affects the rate of future learning about those cues.

Predictiveness principles

The aim of this study was to compare and test the predictions of two attention-based models of associative learning in the inverse base-rate effect when global

outcome frequency was manipulated between groups. Both EXIT and Mackintosh could predict a reduction in the rare choices for conflicting trials when global outcome frequency was matched in the balanced group. However, only EXIT could predict differences in attention to the context between groups. Although EXIT does not provide a perfect account of all the results presented here, on the whole, the results are far more consistent with EXIT than with Mackintosh. Across several conditions using eye-gaze measures at two critical time points within a trial, there was considerable evidence that attention to C over A is stronger than attention to B over A

We therefore showed that two models derived from the same basic principle make opposite predictions about the relative attention paid to a rare predictor versus a common predictor of an outcome. The results demonstrate that the specific formalisation of these models is important, such that very different predictions can be derived based on how their learning algorithms operate.

There is a small but growing list of attention-based learning phenomena that are not consistent with the traditional view of learned predictiveness, as it was originally characterised by Mackintosh (1975). For instance, under some circumstances, associability changes appear to be controlled by absolute predictiveness rather than relative predictivenesss (Kattner, 2015; Le Pelley et al., 2010; Livesey et al., 2011). In other tasks, participants appear to invest more attention after they have encountered strong prediction error (Griffiths, Johnson & Mitchell, 2011; Beesley et al., 2015). There are also questions remaining about whether predictiveness-driven changes operate exclusively on predictive cues or whether similar processes affect outcomes as well (see Griffiths and Thorwart, 2017; Thorwart et al., 2017; Quigley et al., 2018). Our data are also largely inconsistent with the traditional Mackintosh formalisation of the learned predictiveness principle, in that

they demonstrate that learning in the inverse base-rate effect generates stronger attention biases for rare predictive cues than for common predictive cues, and that the predictiveness of the context may be partly instrumental in generating this effect. There are, of course, several attention-based models that may be relevant for understanding attention biases in the inverse base-rate effect; for example, models that reconcile the influence of predictiveness and uncertainty on attention to cues (Le Pelley, 2004; Esber & Haselgrove, 2011). Future theoretical advances will require a variety of new test beds to determine the limits of attentional change as a consequence of predictive learning and the experimental parameters that control it.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Beesley, T., Hanafi, G., Vadillo, M. A., Shanks, D. R., & Livesey, E. J. (2018). Overt attention in contextual cuing of visual search is driven by the attentional set, but not by the predictiveness of distractors. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 44*, 707-721.

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *The Quarterly Journal of Experimental Psychology*, *68*(11), 2175-2199.

Bohil, C. J., Markman, A. B., & Maddox, T. (2005). A feature-salience analogue of the inverse base-rate effect. *The Korean Journal of Thinking & Problem Solving*, *15*, 17-28.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer-Verlag Media.

Cobos, P. L., Vadillo, M. A., Luque, D., & Le Pelley, M. E. (2018). Learned predictiveness acquired through experience prevails over the influence of conflicting verbal instructions in rapid selective attention. *PLoS ONE, 13*, e0200051. DOI: 10.1371/journal.pone.0200051

Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology, 50*, 131-138.

Don, H. J., Goldwater, M. B., Otto, A. R., & Livesey, E. J. (2016). Rule abstraction, model-based choice, and cognitive reflection. *Psychonomic bulletin & review*, *23*, 1615-1623.

Don, H. J. & Livesey, E. J., (2015). Resistance to instructed reversal of the learned predictiveness effect. *The Quarterly Journal of Experimental Psychology, 68*, 1327-1347.

Don, H. J., & Livesey, E. J. (2017). Effects of outcome and trial frequency on the inverse base-rate effect. *Memory & cognition*, *45*, 493-507.

Easdale, L. E., Le Pelley, M. E., & Beesley, T. (2018). *The onset of uncertainty facilitates the learning of new associations by increasing attention to cues.* The Quarterly Journal of Experimental Psychology. Advance online publication. DOI: 10.1080/17470218.2017.1363257

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society of London B: Biological Sciences*, *27*, 2553-2561.

Failing, M., & Theeuwes, J. (2018). Selection history: How reward modulates selectivity of visual attention. *Psychonomic bulletin & review*, *25*, 514-538.

Feldmann-Wüstefeld, T., Uengoer, M., & Schubö, A. (2015). You see what you have learned. Evidence for an interrelation of associative learning and visual selective attention. *Psychophysiology*, *52*, 1483-1497.

Griffiths, O., Johnson, A. M., & Mitchell, C. J. (2011). Negative transfer in human associative learning. *Psychological science*, *22*(9), 1198-1204.

Griffiths, O., & Thorwart, A. (2017). Effects of outcome predictability on human learning. *Frontiers in psychology*, *8*, 511.

Johansen, M. K., Fouquet, N., & Shanks, D. R. (2010). Featural selective attention, exemplar representation, and the inverse base-rate effect. *Psychonomic Bulletin & Review, 17*, 637-643.

Juslin, P., Wennerholm, P., & Winman, A. (2001). High-level reasoning and base-rate use: Do we need cue-competition to explain the inverse base-rate effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 849–871.

Kalish, M. L. (2001). An inverse base rate effect with continuously valued stimuli. *Memory & Cognition, 29*, 4, 587–597.

Kalish, M. L., & Kruschke, J. K. (2000). The role of attention shifts in the categorization of continuous dimensioned stimuli. *Psychological Research, 64*, 105-116.

Kattner, F. (2015). Transfer of absolute and relative predictiveness in human contingency learning. *Learning & Behavior, 43*, 32–43.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*, 671–680.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 3-26.

Kruschke, J. K. (2001a). The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1385-1400.

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

Kruschke, J. K. (2003). Attentional theory is a viable explanation of the inverse base rate effect: A reply to Winman, Wennerholm, and Juslin (2003). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1396-1400.

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

Lamberts, K., & Kent, C. (2007). No evidence for rule-based processing in the inverse base-rate effect. *Memory & Cognition, 35,* 2097–2105.

Larkin, M. J., Aitken, M. R., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24,* 1331.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology Section B, 57,* 193-243.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2011). Overt attention and predictiveness in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37,* 220.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2014). Relative salience versus relative validity: Cue salience influences blocking in human associative learning. *Journal of Experimental Psychology: Animal Learning & Cognition, 40,* 116-132.

Le Pelley, M. E., & McLaren, I. P. L. (2001). Retrospective revaluation in humans: Learning or memory? *The Quarterly Journal of Experimental Psychology: Section B, 54*, 311-352.

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology: B, Comparative and Physiological Psychology, 56*, 68–79.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016, August 8). Attention and associative learning in humans: An integrative review. *Psychological Bulletin 142,* 1111-1140.

Le Pelley, M. E., Turnbull, M. N., Reimers, S. J., & Knipe, R. L. (2010). Learned predictiveness effects following single-cue training in humans. *Learning & Behavior*, *38*, 126-144.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Livesey, E. J., Harris, I. M., & Harris, J. A. (2009). Attentional Changes During Implicit Learning: Signal Validity Protects a Target Stimulus from the Attentional Blink. *Journal of Experimental Psychology: Learning Memory and Cognition, 35*, 408-422.

Livesey, E. J., Thorwart, A., De Fina, N. L., & Harris, J. A. (2011). Comparing learned predictiveness effects within and across compound discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 446-465.

Lochmann, T., & Wills, A. J. (2003). Predictive history in an allergy prediction task. In *Proceedings of EuroCogSci* (Vol. 3, pp. 217-222).

Lovejoy, E. (1968). Attention in discrimination learning: A point of view and a theory. San Francisco, CA: Holden-Day.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66,* 81.

Luque, D., Vadillo, M. A., Gutiérrez-Cobo, M. J., & Le Pelley, M. E. (2018). The blocking effect in associative learning involves learned biases in rapid attentional capture. *The Quarterly Journal of Experimental Psychology, 71,* 522–544.

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

Medin, D . L., & Bettger, J. G., (1991). *Sensitivity to changes in base-rate information. The American Journal of Psychology, 104,* 311-332.

Medin, D. L., & Edelson, S. M., (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 1,* 68-85.

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

O'Bryan, Worthy, D. A., Livesey, E. & Davis, T. (2018). Model-based fMRI reveals dissimilarity processes underlying base rate neglect. *eLife, 7*:e36395. DOI: 10.7554/eLife.36395

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

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32,* 3–25.

Quigley, M. C., Eatherington, C. J., & Haselgrove, M. (2018). Learned Changes in Outcome Associability. The Quarterly Journal of Experimental Psychology, 1-13.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive psychology*, *51*, 1-41.

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

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA 2000*, 71–78. New York, NY: ACM.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Shanks, D. R. (1992) Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science, 4*, 3-18.

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

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E. J., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: a common model for category accentuation and illusory correlation. *Journal of personality and social psychology*, *96*, 305-323.

Shone, L. T., Harris, I. M., & Livesey, E. J. (2015). Automaticity and Cognitive Control in the Learned Predictiveness Effect. *Journal of Experimental Psychology: Animal Learning and Cognition*, 41, 18-31.

Suret, M., & McLaren, I. P. L. (2005). Elemental representation and associability: An integrated model. *New directions in human associative learning*, 155-187.

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

Thorwart, A., Livesey, E. J., Wilhelm, F., Liu, W., & Lachnit, H. (2017). Learned predictiveness and outcome predictability effects are not simply two sides of the same coin. *Journal of Experimental Psychology: Animal Learning and Cognition, 43*, 341-365.

Vanderbilt, D., & Louie, S. G. (1984). A Monte Carlo simulated annealing approach to optimization over continuous variables. *Journal of computational physics, 56*, 259–271.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, *25*, 127-151.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222–236.

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

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

Wills, A. J., Lavric, A., Hemmings, Y., Surrey, E. (2014). Attention, predictive learning, and the inverse base-rate effect: Evidence from event-related potentials. *Neuroimage, 87*, 61-71.

Winman, A., Wennerholm, P. & Juslin, P. (2003). Can attentional theory explain the inverse base rate effect? Comment on Kruschke (2001). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1390-1395.

Winman, A., Wennerholm, P., Juslin, P. & Shanks, D. R. (2005). Evidence for rule-based processes in the inverse base-rate effect. *The Quarterly Journal Of Experimental Psychology, 58A*, 789–815.

Appendix

Mackintosh Model

The original version of the Mackintosh model includes a separable error term for each cue:

$$\Delta V_A = S\alpha_A(\lambda - V_A) \quad \text{(A1)}$$

where ΔV_A is the change in associative strength of cue A, and S and α are learning rate parameters. The error term, $(\lambda - V_A)$, represents the discrepancy between the magnitude of the outcome on that trial (λ) and the extent to which cue A predicts the outcome, or the individual associative strength of cue A (V_A). A small error term indicates that the cue is a good predictor of the outcome, while a large error term indicates that the cue is a poorer predictor. Mackintosh (1975) allows stimulus specific α to change on each trial according to experience of a cue's predictiveness, such that:

$$\Delta\alpha_A > 0 \ \ if \ |\lambda - V_A| \ < \ |\lambda - V_x|$$
$$\Delta\alpha_A < 0 \ \ if \ |\lambda - V_A| \ \geq \ |\lambda - V_x| \quad \text{(A2)}$$

where V_x is the associative strength of all other stimuli present on that trial. If $|\lambda - V_A|$ is less than $|\lambda - V_x|$, α_A will increase as cue A is a better predictor of the outcome than all other stimuli. If $|\lambda - V_A|$ is greater than or equal to $|\lambda - V_x|$, α_A will decrease as cue A is an equal or poorer predictor of the outcome than the other stimuli present.

A separable error term limits the ability of the model to account for several learning phenomena, such as conditioned inhibition. Subsequent variations of Mackintosh (e.g. Le Pelley, 2004; Suret & McLaren, 2005; Pearce & Mackintosh, 2010) instead use a summed error term, which can better capture cue competition or interactions between the predictions of cues (as in the case of conditioned inhibition).

It is possible that a version of Mackintosh with a summed error term may better account for the apparent discrepancy between attention and associative strength. We therefore used a modified version of the Mackintosh model, in which a summed error term is used for the associative strength connecting cue A to outcome k:

$$\Delta V_{Ak} = S\alpha_A(\lambda - \Sigma V) \tag{A3}$$

The following equation was used to implement Equation A2 computationally, after the associative strengths have been updated via Equation A3:

$$\Delta\alpha_A = \theta(\overline{E} - E_A) \times \alpha_A(1 - \alpha_A) \tag{A4}$$

where θ is a parameter that controls the rate of change of α, and E_A represents the total prediction error of cue A summed across all outcomes k:

$$E_A = \sum_{j=1}^{k} |\lambda_j - V_{Aj}| \tag{A5}$$

$\overline{E}$ represents the total prediction error of each cue present on the current trial, summed across all outcomes k, then averaged across all cues C, which provides an estimate of mean predictiveness for the cues present:

$$\overline{E} = \frac{1}{C}\sum_{i=1}^{C}\sum_{j=1}^{k} |\lambda_j - V_{ij}| \tag{A6}$$

Thus a strongly predictive cue should predict the presence of one outcome but also the absence of others. The term $(\overline{E} - E_A)$ reflects how predictive a given cue is relative to all cues present in a way that is analogous to Equation A2. If only two cues are present with no context or common element then $(\overline{E} - E_A)$ will be positive for one and negative for the other unless they are exactly equally predictive, but once context is also considered as an additional cue, this will not necessarily be the case. In

equation 4, $\alpha_A(1 - \alpha_A)$ ensures that α approaches asymptote ($0 < \alpha < 1$), rather than becoming too large or becoming negative.

A cue that strongly inhibits an outcome might end up with a large summed prediction error even when that outcome does not occur, because $|\lambda - V|$ will be positive. Therefore all negative Vs are converted to 0 for the purpose of calculating each of the two error terms $\overline{E}$ and E_A. To determine choice probabilities, we used the same derivative of the Luce (1959) choice rule used in EXIT to map associative strengths to response probabilities:

$$p(c) = e^{\phi \Sigma V_c} / \sum_k e^{\phi \Sigma V_k} \quad \text{(A7)}$$

Parameter description

The EXIT model has seven free parameters:

1. *Exemplar specificity*: the specificity of the exemplar nodes (c in equation 3 in Kruschke (2001a); range: 0.01 – 20.0)
2. *Attention capacity*: or the attention normalisation power (P in equation 5 in Kruschke (2001a); range: 0.1 – 20.0)
3. *Attention shift rate*: A positive constant of proportionality (λ_g in equation 7 in Kruschke (2001a); range: 0.1 – 20.0)
4. *Choice decisiveness*: a response probability scaling constant that converts output activation to response probability (ϕ in equation 2 in Kruschke (2001a); range: 0.1 – 20.0)
5. *Output weight learning rate*: the associative weight learning rate (λ_w in equation 8 in Kruschke (2001a); range: 0.01 – 1.0)
6. *Gain weight learning rate*: the learning rate for the associative weights from exemplar nodes to gain nodes (λ_x in equation 9 in Kruschke (2001a); range:

0.01 – 1.0)

7. *Bias salience*: The salience of the bias (context) cue, (σ in equations 3 and 4 in Kruschke (2001a); range: 0.01 – 2.0. All other cue saliences were fixed at 1.0)

The parameter values were constrained to values between the upper and lower limits shown in brackets. To the best of our knowledge, Kruschke (2001a) did not include limits on the parameters in his model simulations. As such, where these parameters typically vary between 0 and 1, e.g. learning rate, these limits were imposed, and where they do not, an arbitrary limit of 20 was imposed.

The implementation of Mackintosh used here has five free parameters:

1. *Learning rate:* (S in Equation A3; range: 0.01 – 1.0)
2. *Initial alpha for cues*: The initial associability of stimuli (range: 0.1 – 1.0).
3. *Initial alpha for context*: The initial associability of the context (range: 0.1 – 1.0).
4. *Theta*: A parameter that controls the rate of change of α (range: 0.1 – 1.0).
5. *Choice decisiveness*: A response probability scaling constant that converts output activation to response probability (ϕ in equation A7; range: 0.1 – 20.0).

Parameter search

Simulated annealing was used to find the best fitting parameters for each model (Kirkpatrick, Gelatt & Vecchi, 1983; Vanderbilt & Louie, 1984), which was run using the simulannealbnd function in the Global Optimisation Toolbox for Matlab. Simulated annealing is a preferred parameter estimation technique for complex models, as it allows upward movements on the error surface, which is useful for avoiding local minima (Lewandowsky & Farrell, 2011). That is, it allows the parameter search to jump out of local minima in order to better find the global minimum. Because the trial randomisation leads to differences in the model

predictions from one run to the next, each simulation was run with 30 simulated participants per group, and the parameter search was repeated five times. The set of parameters with the lowest root-mean-square error (RMSE; the average difference between the predicted and observed values) at the group level were selected. The best fitting parameters for EXIT were: *exemplar specificity* = 15.26; *attention capacity* = 14.78; *choice decisiveness* = 6.17; *attention shift rate* = 17.29; *output weight learning rate* = 0.37; *gain weight learning rate* = 0.09; *bias salience* = 0.02; *RMSE* = 5.81. The best fitting parameters for Mackintosh were: *S* = 0.32; *initial alpha for cues* = 0.14; *initial alpha for context* = 0.1; *theta* = 0.99; *choice decisiveness* = 8.88; *RMSE* = 5.72. These parameters were then used to estimate the model predictions for the test trials. This was again run five times, and the simulated data with the lowest RMSE are reported.