Attention biases in the inverse base-rate effect: Prediction error or novelty?

Shu Chen (sche5879@uni.sydney.edu.au)
School of Psychology, Brennan Maccallum Building, A18 Manning Rd,
Camperdown, NSW, Australia

Hilary J. Don (h.don@ucl.ac.uk)
Division of Psychology and Language Sciences, 26 Bedford Way,
London WC1H 0AP, United Kingdom

Evan J. Livesey (evan.livesey@sydney.edu.au)
School of Psychology, Brennan Maccallum Building, A18 Manning Rd,
Camperdown, NSW, Australia

Abstract

Attention-based models of categorization and associative learning have received considerable support from human learning phenomena in which multiple predictive cues compete for association with outcomes. Among these, several phenomena (e.g. the highlighting effect and inverse base-rate effect) lend strong support to models that propose attention is driven by the experience of prediction error, and is distributed strategically to minimize prediction error during current and future learning. Here we explore the possibility that attention is determined instead by a relatively simple combination of stimulus novelty and association strength. We apply the model to several key findings in the literature on the inverse base-rate effect and related phenomena. Overall, the model provides a surprisingly good account of complex behavioral biases.

Keywords: learning, attention, categorization, prediction error, inverse base-rate effect, novelty

Prediction error and attention

Extracting relevant predictive information from our surroundings is a fundamental cognitive competency. It allows us to draw inferences about cause and effect from a rich and complex environment and to prioritise the processing of the signals that are most likely to aid us in achieving our goals. Several influential models of category learning (Kruschke, 2001) and associative learning (Mackintosh, 1975) have proposed that we shift the balance of our attention to competing cues in our environment when we encounter *prediction error*, that is, the discrepancy between predicted and observed events. Put simply, these models suggest that we direct attention away from sources of large prediction errors (e.g. cues that are associated with an outcome that doesn't occur, or features that predict an incorrect category label) and towards signals that minimize prediction error (e.g. cues that are associated with the outcome that eventuates, or features that predict the correct category label). This prioritisation of attention could serve an important adaptive function, allowing us to learn predictive relationships faster and devote more resources to the features of our environment that are actually useful. Under certain circumstances, it also leads to some surprising predictions about the decisions a learner will make when faced with conflicting or ambiguous predictive information. Evidence in the form of behavioral results that support these predictions is thus important for our theoretical understanding of attention and its role in learning.

The inverse base-rate effect

The inverse base-rate effect (IBRE) is one such example in human learning, where learners exhibit a bias toward predicting the lower frequency of two predicted outcomes when faced with conflicting information. A typical IBRE design is depicted in Table 1, and is usually presented within an explicit predictive learning task, for instance where the participant must diagnose patient illnesses (outcomes o1-o4) on the basis of presented symptoms (cues A-L) (e.g. see Medin & Edelson, 1988). Note that these tasks usually use duplicates of a simple design—the first two rows of Table 1—with three potentially predictive cues (thus we will refer only to cues A, B, and C, as one instance of this design).

Table 1: An example design of an IBRE experiment.

Training	Test Trials	Typical finding
3× AB–o1	A; ABC	o1 > o2
1× AC–o2	**BC**	**o2 > o1**
3× DE–o1	D; DEF	o1 > o2
1× DF–o2	**EF**	**o2 > o1**
3× GH–o3	G; GHI	o3 > o4
1× GI–o4	**HI**	**o4 > o3**
3× JK–o3	J; JKL	o3 > o4
1× JL–o4	**KL**	**o4 > o3**

Note: Letters A-L refer to predictive cues or diagnostic features, o1-o4 refer to outcomes or category labels that the participant is typically asked to predict and learn about through trial and error. Only the critical test trials are shown. The typical IBRE result for conflicting test trials is highlighted in bold.

Learners are exposed to two types of trials (common and rare) during training, with common trial types occurring more frequently (e.g. three times as often) than the rare trial type. On each common trial, a predictive cue B combined with a less predictive cue A always leads to a disease outcome (o1). On each rare trial, another predictive cue C combined with

the same less predictive cue A consistently causes a second disease outcome (o2). Since A is associated with both o1 and o2, it is an imperfect predictor of the outcome in comparison to the perfect predictors B and C that accompany it.

At test, learners are asked to predict the outcome arising from the simultaneous presentation of the two distinctive cues B and C. The conflict lies in the different outcome that B and C respectively elicits, each usually with close-to-ceiling accuracy during training. The conditional probabilities of the outcomes—p(o1|B) and p(o2|C)—are equivalent and effectively cancel each other out. However, it seems reasonable to assume that learners should favor the common outcome o1 as the more likely event for its more frequent overall occurrence (e.g. more patients are observed to suffer from disease o1 than o2), and indeed given a summary of the statistics of the task, people tend to do just this (Don, Worth & Livesey, 2021). However, when presenting the cue-outcome relationships in sequential fashion in a predictive learning task, researchers have repeatedly found an overprediction for the infrequent o2 on the conflicting test trial, which has been referred to as the IBRE (see Don et al., 2021 for a review).

This rare choice bias is specific to the BC conflicting trials. Other ambiguous test trial types such as presenting the imperfect predictor A by itself or presenting a *combined* test trial with all three cues ABC tend to elicit more common outcome predictions. Not only does the rare outcome bias on BC trials go against the normative use of statistical regularities, it is also challenging to traditional learning models such as the influential Rescorla-Wagner model (1972; see Markman, 1989).

Similar rare choice bias has been revealed in a complementary phenomenon called highlighting (Kruschke, 1996). In a typical highlighting design, learners first observe that a pair of cues (AB) lead to an outcome (o1), and subsequently observe that one of the cues in the compound (A), combined with another cue (C), leads to a different outcome (o2). As a result, the association between the distinctive cue C and later-learned o2 is usually augmented or highlighted, biasing learners to the rare outcome in the face of ambiguous BC test compound. Highlighting and the IBRE are assumed to be the result of the same learning and decision processes, and both are well accounted for by theories that assume changes in attention are governed by the experience of prediction error (Don et al., 2021).

Attention-based models of learning

While many associative models fall short in explaining the IBRE, this effect can be accommodated by attentional learning theories that conceive of learning as being determined by learned attention to cues possessing better predictive value as well as learned inattention to cues with poor predictive utility. It is widely assumed by this class of models that learners selectively attend to cues that are better predictors of the outcome relative to other concomitant cues and the key determinant of these changes is prediction error (Mackintosh, 1975).

Perhaps the most widely and successfully applied explanation for the IBRE is provided by the EXIT model, which is a key example of the application of prediction error as a driver of attentional change (Kruschke, 2001a; Kruschke, 2001b). EXIT is a connectionist model where each constituent cue of a compound is represented by an input node and each outcome is represented as an output node. Learning of a cue-outcome relationship is expressed as building associative connections between the relevant input and output nodes. EXIT also assumes that learning is modulated by the amount of attentional resources devoted to a cue. EXIT makes the further assumption that attention increases to good predictors at a cost to attention paid to poor predictors due to the limited total attention capacity. In the network representation, each cue node is connected to an exemplar node which is activated to the extent that the cue is similar to the exemplar cue combination. The exemplar node is connected to the gain node where attention is normalised to determine the final attention value. The gain node then spreads activation to the outcome node to make a prediction based on the activation of the outcome.

Importantly, the model recruits error-driven attention shifting as the major mechanism for cue competition effects. In the context of the IBRE, attention rapidly shifts to the more predictive cues B and C and away from the less predictive shared cue A after experiencing error on each trial. As there are more frequent encounters with AB trials, A becomes more predictive of o1 than o2. Attending to A on AC trials elicits the prediction of o1 which is discrepant with the observed o2, attention therefore shifts away from the error-causing A toward the perfectly predictive C in response to this large error, leading to a particularly strong C-o2 association. The learned attentional shift to C is then also transferred to the BC test trial further tipping choices towards a preference for the outcome associated with C (i.e. o2). It should be noted here that the rare attentional bias may be captured by a simpler account offered by Le Pelley et al. (2016) where the amount of attention allocated to a cue is assumed to be directly proportional to its associative strength. If C accrues greater associative strength than B and the context (and their combination), then more attention will be attracted to C on the BC test trial.

Evidence for prediction error driven attention

Attentional bias driven by prediction error has been considered a likely source of the IBRE due to several findings that seem particularly consistent with this explanation. For instance, the presence of the shared cue A has been shown to be a critical condition for the observation of the rare choice bias. This imperfect predictor allows re-direction of attention towards the more predictive component of the rare compound thereby creating more opportunities for the rare predictor to be learned (Wills et al., 2014).

Recent empirical evidence has provided direct support for EXIT's prediction of the rare attentional bias. Don et al. (2019) found that attention as indicated by overt eye gaze on the rare predictor relative to the imperfect predictor was

considerably longer than on the common predictor relative to the imperfect predictor during both the prediction phase and the feedback phase (see Kruschke et al., 2005 for a similar result for highlighting). Further tests of associability difference in subsequent learning revealed a persistent attention benefit for the rare predictor (Don & Livesey, 2021). A simplified version of the design presented by Don & Livesey (2021) is shown in Table 2. The logic of this design rests on the assumption that cues which have gained attention during stage 1 of training will carry over some of that attentional advantage to stage 2 and will thus be learned about faster. To test this, cues that were initially trained as common predictors and rare predictors of o1 and o2 in stage 1 were then rearranged to predict two new outcomes in stage 2. The cues were combined so that a previous common predictor was always competing for attention with a previous rare predictor.

Two types of test trial then followed. On summation test trials, two cues with the same stage 1 role and predicting the same stage 2 outcome (e.g. in BI, B and I were both common predictors in stage 1 and both predicted o3 in stage 2) were combined. The key summation result was that accuracy for correct stage 2 outcomes was higher for stage 1 rare predictors than for stage 1 common predictors.

On negation test trials, a common predictor from stage 1 predicting one outcome in stage 2 was always combined with a rare predictor from stage 1 predicting the *other* outcome in stage 2 (e.g. in BC, B was a common predictor in stage 1 and predicted o3 in stage 2, whereas C was a rare predictor in stage 1 and predicted o4 in stage 2). The key negation result was that participants tended to choose whichever stage 2 outcome was predicted by the stage 1 rare predictor (e.g. choosing o4 over o3 when presented with BC).

Table 2: An example design for cue attention transfer following IBRE training

Training 1	Training 2	Test	Finding
3× AB–o1	BI–o3	BE	Summation
1× AC–o2	CH–o4	HK	% o3: IL > BE
3× DE–o1	EL–o3	CF	% o4: CF > HK
1× DF–o2	FK–o4	IL	Negation
3× GH–o1		BC	o4 > o3
1× GI–o2		EF	o4 > o3
3× JK–o1		HI	o3 > o4
1× JL–o2		LK	o3 > o4

Note: Letters A-L refer to predictive cues or diagnostic features, o1-o4 refer to outcomes or category labels.

Both of these results suggest that more was learned about the stage 1 rare predictors in stage 2, despite the fact that all cues serve the same objective relationship with the respective outcomes in stage 2 and occurred an equal number of times in stage 2. The summation and negation effects can only be achieved by virtue of the roles that the cues played in stage 1 and strongly suggests greater attention paid to rare predictors. The result was particularly strong after relatively brief stage 1 training (see Don & Livesey, 2021).

One of the major criticisms of the EXIT model is its relative complexity compared to other models of learning based on similar principles (e.g. see Paskewitz & Jones, 2020). It characterizes attention shifting as a strategic and highly dynamic process that occurs immediately on encountering prediction error. Although strong evidence exists that attention is correlated with the predictiveness of cues, there are simpler ways in which learning about a predictive relationship might guide attention. For instance, Le Pelley et al. (2016) suggested that attention paid to a cue may simply be proportional to the strength of the associations that the cue has developed with task relevant outcomes. However, it has been suggested that one of the features of the IBRE may provide evidence in favor of relatively complex mechanisms proposed in the EXIT model.

Paradoxically, when each cue is tested by itself, accuracy for B has been reported in several papers to be higher than accuracy for C (e.g. Wills et al., 2014; Inkster et al., 2022b). This difference is suggestive of a stronger association between the common predictor and the common outcome than between the rare predictor and the rare outcome, which is at odds with the rare bias observed on the BC test trial. This is difficult to reconcile with simpler models of attention and suggests that attention to the rare predictor may be a specific consequence of experiencing strong prediction error on those trials, rather than simply a product of the associative strength of the rare predictor itself.

Prediction error or novelty?

These findings have been plausibly explained by appealing to attention being driven by the larger prediction error encountered on rare trials. However, as the rare predictor is experienced less often than the common predictor, the rarity inherent in its infrequent presentation may suffice to attract more attention than does the common predictor when the two are encountered during IBRE training and also when presented together during subsequent transfer tests. While the larger prediction error on rare trials than on common trials appears to be the most plausible and widely accepted explanation for the IBRE (Don & Livesey, 2021), an alternative explanation in terms of the inherent novelty of the rare predictor remains unexplored.

Novelty and prediction error are two closely connected yet distinct concepts in psychology (Barto et al., 2013). An experience is usually considered novel if it is new and has the quality of not being encountered before. Prediction error on the other hand refers to the mismatch between prediction and actuality, which requires a comparison to be made between the expectation evoked prior to the occurrence of an event and the actual event observed or experienced. Here we use the term novelty to refer to the extent to which an event is unexpected. We rely on the assumption that, within the context of the experiment, cues that occur more often also come to be expected more strongly on future trials as a consequence of learning the base-rates of their occurrence.

We operationalise this conception of cue novelty by assuming that an association forms between the context and each cue, such that cues come to be predicted by the experimental context and the strength of that prediction downweights attentional gain to that cue. This conception of novelty is a convenience and we acknowledge that other non-associative mechanisms could be used to govern changes in cue familiarity. Changes in context-cue associations are largely independent of the formation of cue-outcome associations on which the simulated response probabilities are based, except that attention to the cues is weighted based on their novelty and the context itself enters into associations with both the cues *and* the outcomes.

The rare predictor is present on fewer occasions than the common predictor and this relative rarity has the potential to attract more attentional resources to it when later combined with the common predictor, leading to the same prediction of the rare outcome on the conflicting test trial as the error-driven account. Novelty-attracted attention has long been reported in developmental literature as preferential looking and longer fixation time for novel stimuli compared to old or repeated ones in human infants (e.g. Franz, 1964; Cohen & Gelber, 1975). Likewise, the phenomenon of latent inhibition, extensively studied in classical conditioning (Lubow & Moore, 1959; see Holmes & Harris, 2010 for a review) and in a range of human learning paradigms (e.g. Forrest, Mather & Harris, 2018; Quinn, Livesey & Colagiuri, 2017), suggests that passive pre-exposure to a cue in the absence of reinforcement could hinder its ability to later enter into meaningful associations by reducing its ability to capture attention. It is worth-noting that novelty alone does not suffice as a satisfactory account for the IBRE (Wills et al., 2014; Don & Livesey, 2017), yet novelty, which we define as how strongly a cue is predicted by the context, combined with the associative strength of a cue, may still provide a relatively simple alternative explanation for some aspects of the IBRE including the rare attentional advantage and the necessity of the shared cue.

Attention as a function of novelty and association.

Here we describe the development of a model in which the allocation of attention is determined by a combination of stimulus novelty (how unexpected the cue is in the context of the experiment) and maximal prior learning about the cue (the strength of the strongest association that the cue holds with any of the relevant outcomes).

The model uses several of the equations defined by Kruschke (2001b) in developing and implementing the EXIT model. For instance, we use the same algorithm to generate associative predictions for each outcome based on attention-weighted associations between cues and outcomes (equation 1 in Kruschke, 2001b), we use the softmax rule to compute prediction probabilities (equation 2 in Krushcke, 2001b), and we use the same delta rule algorithm to update cue-outcome weights (equation 8 in Krushcke, 2001b). However here the attention to each cue i is determined by a combination of 1) cue novelty (nov_i), and 2) the maximum association strength linking the cue with an outcome ($maxW_i$). To determine nov_i associative weights are allowed to develop between the context (originally represented in EXIT as an additional "bias" cue node) and each of the cues. The weight linking the context to cue i—$w_{ctxt,i}$ —starts at zero and is updated at the end of each trial via a simple delta algorithm, as follows:

$$\Delta w_{ctxt,i} = S(obs_i - w_{ctxt,i})$$

Where S is a learning rate parameter and obs_i is a teaching signal that simply takes a value of 1 if the cue is present and 0 if the cue is absent. The value of $w_{ctxt,i}$ causes an exponential decrease in the gain for the cue. That is, as the learner develops an expectation of seeing the cue in the context of the learning experiment, its ability to capture attention begins to reduce. This is achieved via the following:

$$nov_i = Ce^{-Cw_{ctxt,i}}$$

Here, C is a parameter affecting the rate of this decline and the relative salience of surprising cues (high values lead to greater initial salience and a faster decline).

The second component determining gain is the absolute magnitude of the strongest association cue i holds with one of the task-relevant outcomes j—$max_j w_{ij}$ —which contributes positively to the gain score according to the following:

$$max_i = 1 + D \max_j w_{ij}$$

Here, D is a scaling parameter, while the addition of 1 merely ensures that novel cues which have no existing association are never impossible to attend to or learn about (following the assumption that novelty carries attentional biases most strongly in early stages of learning). These two determinants of gain are then simply multiplied:

$$gain_i = nov_i max_i$$

Finally, we apply the same limited capacity attention algorithm to transform these gain values into an attention score for each cue, dependent on which cues are present on each trial (equation 5 in Kruschke, 2001b).

Simulations of key IBRE results

As an initial test of the capabilities of the model we ran simulations of the designs shown in tables 1 and 2, as well as the effect of the common and rare training trials sharing an imperfect predictor (i.e., shared cue condition: AB-o1 / AC-o2 versus unique cue condition: AB-o1 / XC-o2). The aim was to demonstrate 1) that the model can reproduce the well-replicated pattern of choice data observed for critical ambiguous trial types associated with the IBRE, 2) that the model also anticipates higher accuracy for common predictors than rare predictors when they are presented individually, 3) that it predicts that the IBRE will be dependent on the shared imperfect predictor (A) across common and rare training trials, and 4) that the model

anticipates attentional transfer that favors rare predictors over common predictors.

For each of the experimental designs, we generated 100 unique trial sequences using randomization of trial order in individual blocks of trials as shown in the table. We ran 1000 simulated participants using these sequences, choosing parameter values from a uniform distribution within parameter bounds described in Table 3.

Table 3: List of free parameters used in simulations.

parameter	description	range
L	Cue-outcome learning rate	.01–1.0
S	Context-cue learning rate	.01–1.0
B	Salience of context (bias)	.05–1.0
P	Attention capacity	1.0–20
T	Choice decisiveness	1.0–20
C	Cue novelty scaling	10.0–20.0
D	Cue association scaling	0.1–2.0

The results of each of these simulations is illustrated across figures 1-4. In each figure, rare choice proportion reflects the probability of choosing the relevant rare outcome divided by the sum of the choice probabilities for the relevant common outcome and relevant rare outcome. For a BC trial this would be p(o2) / [p(o1)+p(o2)]. For trials where the associated outcome should be unambiguous, accuracy refers to the probability of choosing the outcome that was previously associated with the cue or cues presented.

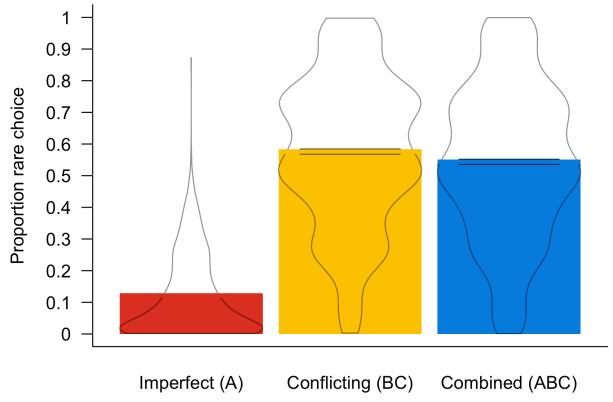

Figure 1: Results of 1000 simulated participants run through the design shown in Table 1 (IBRE).

Figure 1 presents the critical ambiguous test trials typically reported in IBRE research. With randomly chosen parameter values, the model successfully replicates the predominant common-outcome choice for the imperfect predictor A, and a consistent rare choice bias for conflicting BC trials. This result replicates the IBRE, though with randomly chosen parameters the effect is fairly modest, with the predicted proportion rare choice often between 0.5 and 0.7. Predictions for combined ABC trials are less consistent, though this reflects similar inconsistencies in empirical findings (see Don et al., 2021).

The model is thus able to reproduce both the ordinal pattern of choice preferences typically observed for the key ambiguous test trials used in the IBRE, as well as the IBRE itself (that is, higher proportion of rare outcome than common outcome choice on conflicting trials). It should be noted that a similar pattern is observed when simulating a highlighting design in which rare (i.e., late) training trials are only introduced after an initial phase comprising only common (i.e., early) trials.

As is illustrated in Figure 2, the model makes these predictions while also yielding higher choice accuracy for common predictors tested alone than for rare predictors tested alone. The model achieves this because the context generally accrues stronger associations with the common outcome than the rare outcome, and therefore the high accuracy for the common predictor is partly a reflection of a context-driven bias towards choosing the common outcome. This explanation is consistent with a possibility raised previously by Le Pelley et al. (2016). Some recent evidence (Inkster et al., 2022b) found that attempts to change the context after initial training have no effect on the IBRE or the higher accuracy for B than C and thus this explanation is one that needs further empirical attention.

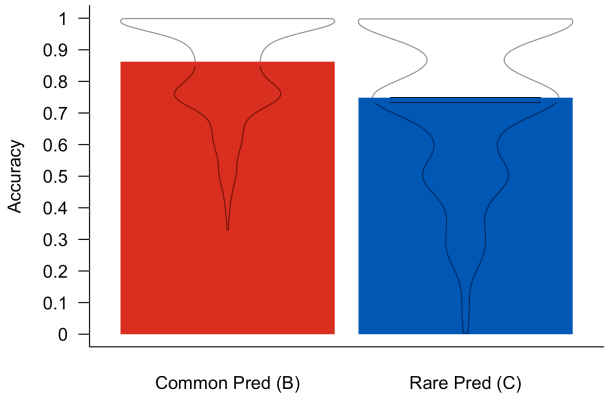

Figure 2: Simulated results comparing accuracy for common and rare predictors when presented in isolation, using the training design outlined in Table 1.

Figure 3 shows the effect of replacing the shared imperfect predictor with cues unique to common and rare training trials. For simulation of the importance of the shared cue, we retained the regular design in Table 1 for two of the sets of overlapping contingencies (e.g. AB-o1/AC-o2, DE-o1/DF-o2) but replaced the shared cue with two unique cues for the other two sets (e.g. GH-o3/MI-o4, NK-o3/JL-o4).

Even though attention is not driven by prediction error in the model, the fact learning of the cue-outcome associations is based on a simple prediction error algorithm means that cues compete for association (they "overshadow" each other) and a bias in attention towards a cue leads to a greater share of the learning being attributed to that cue. Since relatively novel cues gain more attention, more is learned about C than A on AC-o2 trials and this bias is stronger than learning for B versus A on AB-o1 trials. In contrast, when there is no

shared cue, C and X are equally novel and have the same objective relationship with o2, meaning they develop associations of comparable strength. In the absence of a strong association between the rare predictor and the rare outcome, choice tends to favor the common outcome or no preference for either common or rare outcome. The presence of the common cue is thus still critical for generating strong learning to the rare predictor C.

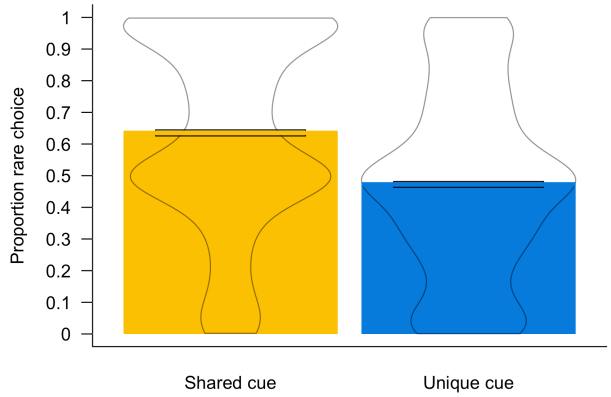

Figure 3: Simulated results comparing conflicting trials when the common and rare predictors were trained with a common versus unique added cue.

Finally, Figure 4 shows the effect of initial IBRE training on later learning about novel outcomes. Here, attentional biases towards the rare predictors, initially because they are more novel and then in addition because they carry a stronger association strength with the rare outcome, transfer to stage 2 learning. This means that the model predicts more accurate responses to summation test trials comprising stage 1 rare predictors compared to those comprising stage 1 common predictors, and for negation trials, predicts that the outcome associated with the stage 1 rare predictors is more likely to be chosen.

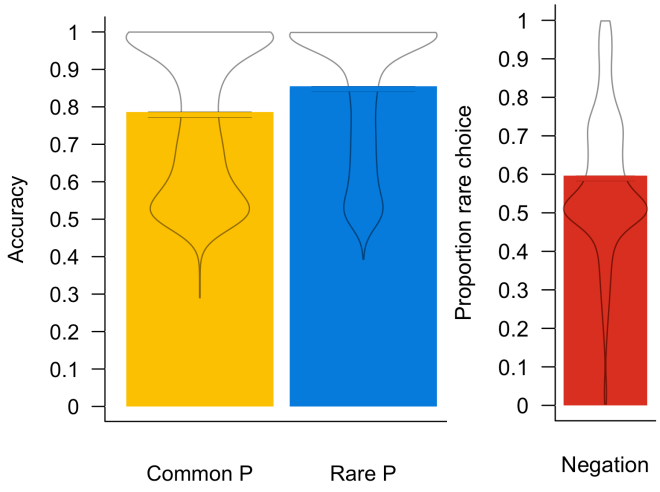

Figure 4: Simulated results for accuracy on summation test trials and rare choice proportion on negation test trials

Discussion

Although there is clear evidence that attention changes as a consequence of learning predictive relationships, debate continues about the factors that determine how and when these changes occur. Attention shifts in response to encountering prediction error is a popular mechanism advocated to explain a range of effects related to human associative learning and categorization. However, the models that have been most successful in implementing this idea are complex, and debate about their relevance has been hampered by a lack of credible alternative explanations.

Here we explored the possibility that at least some of these phenomena may be explained by differences in cue novelty rather than the prediction error encountered in the presence of those cues. There is strong evidence from multiple lines of research that indicates processing (including attentional selection) changes as stimuli become more familiar. To the extent that a simple factor like cue novelty can explain a widely replicated learning phenomenon like the IBRE, that explanation should be taken seriously.

Here we applied a basic prediction error learning algorithm in which the learning rate for a particular cue was determined by two factors, 1) the extent to which that cue was predicted by the context, and 2) the strength of the maximal association between that cue and any task relevant outcome. The model provides a surprisingly good fit to data that have been held up as clear evidence for attention shifting models. By including the strength of the maximal association as a determinant of attention, the model should in principle be able to account for other attention-based biases in learning (e.g. Don & Livesey, 2015; Le Pelley & McLaren, 2003; Livesey & McLaren, 2007) but future work will be needed to determine the limits of its capabilities in these and other domains. It should be noted that the model in its current form relies heavily on context associations. The possible roles of context learning in the IBRE are a matter of ongoing debate (e.g. Don & Livesey, 2017; Don et al., 2019; Inkster et al., 2022b), and more research is needed to test whether this component of our explanation is plausible.

Some findings in the IBRE literature are clearly beyond the scope of category learning models of this variety, regardless of whether attention is driven by prediction error or other factors, including several that provide tentative support for inferential reasoning accounts (see Don et al., 2021). There is also neural evidence from imaging studies consistent with conflicting explanations for the choice of the rare outcome (Inkster et al., 2022a; O'Bryan et al., 2018). However, the evidence that attention biases are involved in the effect is now fairly compelling. Further research is required to ascertain the source of those attention biases and the role they play in producing choices that seem to defy normative reasoning based on the frequency of events.

References

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or Surprise? *Frontiers in Psychology, 4*, 1-15.

Cohen, L. B., & Gelber, E. R. (1975). Infant visual memory. In L. B. Cohen & P. Salapatek (Eds.), *Infant perception: From sensation to cognition.* New York: Academic Press.

Don, H. J. & Livesey, E. J. (2015). Resistance to instructed reversal of the learned predictiveness effect. *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.

Don, H. J. & Livesey, E. J. (2021). Attention biases in the inverse base-rate effect persist into new learning. *Quarterly Journal of Experimental Psychology, 74*, 669-681.

Don, H. J., Beesley, T., & Livesey, E. J. (2019). Learned predictiveness models predict opposite attention biases in the inverse base-rate effect. *Journal of Experimental Psychology: Animal Learning & Cognition, 45*, 143-162.

Don, H. J., Worthy, D. A., & Livesey, E. J. (2021). Hearing hooves, thinking zebras: A review of the inverse base-rate effect. *Psychonomic Bulletin & Review, 28*, 1142–1163.

Fantz, R. (1964). Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones. *Science, 146*(3644), 668–670.

Forrest, D. R., Mather, M., & Harris, J. A. (2018). Unmasking latent inhibition in humans. *Quarterly Journal of Experimental Psychology, 71*, 380-395.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. Journal of Experimental Psychology: General, 117, 227–247.

Holmes, N., & Harris, J. (2010). Latent inhibition. In C. J. Mitchell & M. E. Le Pelley (Eds.), Attention and Associative Learning: From Brain to Behaviour (pp. 99-130). Oxford: Oxford University Press.

Inkster, A.B., Milton, F., Edmunds, C.E.R., Benattayallah, A., & Wills, A.J. (2022a). Neural correlates of the inverse base-rate effect. *Human Brain Mapping, 43*, 1370-1380.

Inkster, A.B., Mitchell, C. J., Schlegelmilch, R., & Wills, A. J. (2022b). Effect of a context shift on the inverse base rate effect. *Open Journal of Experimental Psychology and Neuroscience, 1*, 22-29.

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. (2005). Learning involves attention. In G. Houghton (Ed.), *Connectionist models in cognitive psychology* (pp. 113–140). Hove, East Sussex, UK: Psychology Press.

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.

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

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

Livesey, E. J. & McLaren, I. P. L. (2007). Elemental Associability Changes in Human Discrimination Learning. *Journal of Experimental Psychology: Animal Behavior Processes, 33*, 148-159.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: The effect of nonreinforced pre- exposure to the conditional stimulus. *Journal of Comparative and Physiological Psychology, 52*(4), 415–419.

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

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General, 118*, 417-421.

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.

O'Bryan, S., Worthy, D., Livesey, E. J., Davis, T. (2018). Model-based fMRI Reveals Dissimilarity Processes Underlying Base Rate Neglect. *eLife, 7*:e36395.

Paskewitz, S., & Jones, M. (2020). Dissecting EXIT. *Journal of Mathematical Psychology, 97*, 102371.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532–552.

Quinn, V. F., Livesey, E. J. & Colagiuri, B. (2017). Latent inhibition reduces nocebo nausea, even without deception. *Annals of Behavioral Medicine, 51*, 432-441.

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.