Attention biases in the inverse base-rate effect persist into new learning

Journal:	*Quarterly Journal of Experimental Psychology*
Manuscript ID	QJE-STD-20-253.R1
Manuscript Type:	Standard Article
Date Submitted by the Author:	10-Aug-2020
Complete List of Authors:	Don, Hilary; The University of Sydney, School of Psychology Livesey, Evan; University of Sydney, School of Psychology
Keywords:	inverse base-rate effect, attention, associability, learned predictiveness, associative learning

DOI: 10.1177/1747021820985522

Attention biases in the inverse base-rate effect persist into new learning.

Hilary J. Don

Evan J. Livesey

The University of Sydney

Please address correspondence to:

Hilary J Don

School of Psychology

The University of Sydney

Sydney, NSW, 2006, Australia

hdon7006@uni.sydney.edu.au

DOI: 10.1177/1747021820985522

Abstract

The inverse base-rate effect is a tendency to predict the rarer of two outcomes when presented with cues that make conflicting predictions. Attention-based accounts of the effect appeal to prioritised attention to predictors of rare outcomes. Changes in the processing of these cues are predicted to increase the rate at which they are learned about in the future (i.e. their *associability)*. Our previous work has shown that the development of the inverse base-rate effect is accompanied by greater overt attention to the rare predictor while participants made predictions, and during feedback, and these biases changed in different ways depending on the stage of training and global base-rate differences. It is unknown whether these gaze patterns reflect the manner in which cues are prioritised for learning or are merely a consequence of learning what the cues predict. This study tested whether the associability of common and rare predictors differed, and if so, how this difference changed as a function of training length and the presence of base-rate differences in the outcomes. Experiment 1 tested cue associability using a second learning task presented after either short or long training. The results suggest an associability advantage for rare predictors that *weakens* with extended training, and is not strongly affected by the presence of global base-rate differences. However, Experiment 2 showed a clear effect of base-rate differences on choice after very brief training, indicating that attention biases as measured by associability change are not sufficient to produce the inverse base-rate effect.

Keywords: Inverse base-rate effect; attention; associability; learned predictiveness, associative learning

DOI: 10.1177/1747021820985522

A wealth of evidence has accumulated over the last 15 years on the way predictive learning influences attention. One of the most consistent findings is that, as a consequence of learning, more predictive cues—those which are particularly useful for predicting a task-relevant outcome—come to command more attention than less predictive cues (Mackintosh, 1975), a phenomenon that is known as the learned predictiveness effect (Le Pelley & McLaren, 2003; Lochmann & Wills, 2003). While this relationship has been shown using a range of attentional measures (see Le Pelley, Mitchell, Beesley, George, & Wills, 2016, for a review), one of the most frequently used and arguably the most directly relevant for the study of learning, is cue *associability*. Cue associability refers to the ease with which a cue can be associated with an outcome in subsequent learning, especially *de novo* learning about a novel cue-outcome relationship. Demonstrations of transfer show that previously predictive cues are learned about more readily than previously non-predictive cues in a new training phase, an effect that has been replicated many times (e.g., Don & Livesey, 2015; Easedale, Le Pelley & Beesley, 2019; Le Pelley & McLaren, 2003; Le Pelley et al., 2011; Livesey, Don, Uengoer & Thorwart, 2019; Livesey & McLaren, 2007; Livesey, Thorwart, De Fina & Harris, 2011; Shone, Harris & Livesey, 2015; Mitchell, Griffiths, Seeto, & Lovibond, 2012). Such is the strength and ubiquity of this learned predictiveness associability effect that progress in honing attention-based theories of learning will arguably require designs that go beyond comparing the associability of a predictive and non-predictive cue. In this respect, Le Pelley et al. (2016) singled out the inverse base-rate effect as a potentially important phenomenon for distinguishing between different models that all anticipate the learned predictiveness effect. The associability changes that accompany the inverse base-rate effect are the focus of the current study.

DOI: 10.1177/1747021820985522

The inverse base-rate effect refers to a seemingly irrational bias in human decision-making (Medin & Edelson, 1988). In demonstrations of this effect, a compound of two cues, AB, predicts outcome 1 (O1), and compound AC predicts outcome 2 (O2), however AB-O1 occurs more frequently than AC-O2. Thus, Cue A is an *imperfect* predictor, as it is paired with both outcomes. Cue B (the *common predictor*) is a perfect predictor of the common outcome, O1 and cue C (the *rare predictor*) is a perfect predictor of the *rare* outcome, O2. After learning these contingencies, participants are given a test phase in which they are presented with 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. 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). It is this choice of the rare outcome given conflicting predictive information that is referred to as the inverse base-rate effect, which has been reliably replicated across different tasks (Dennis & Kruschke, 1998; Johansen, Fouquet & Shanks, 2010; Kalish, 2001; Kalish & Kruschke, 2000; Kruschke, Kappenman & Hetrick, 2005; Lamberts & Kent, 2007; Sherman et al., 2009; Wills, Lavric, Hemmings & Surrey, 2014).

Typical explanations of the inverse base-rate effect rely on prioritised attention to cue C during training (Kruschke, 1996; 2001a). There are of course competing explanations for the effect (e.g. O'Bryan et al., 2018), however the current paper will focus primarily on these

DOI: 10.1177/1747021820985522

attentional accounts. Due to the relative frequency of AB-O1 trials, the association between both cues and O1 is learned well. On rare AC trials, A elicits an incorrect prediction of O1. In order to reduce error and preserve learning about AB-O1 trials, attention shifts away from the imperfect predictor towards the more predictive cue C. Due to this increase in attention, the association between C and O2 is stronger than the association between B and O1, such that BC trials elicit an O2 response as a result of simple associative strength. In addition, prioritised attention to C may transfer to BC trials, such that C is more likely to control responding. This account has been formalized in Kruschke's EXIT model (2001b), which is based on learned predictiveness principles like those proposed by Mackintosh (1975). Yet the EXIT model and variants of Mackintosh's model have been shown to make different predictions regarding attention to cues in the inverse base-rate effect (Don, Beesley & Livesey, 2019). Although EXIT is a relatively complex model containing several mechanisms, Paskewitz and Jones (2020) have shown that the EXIT model only requires rapid attentional shifts or attentional competition components in order to explain most experimental effects.[1]

Patterns of gaze biases support the idea that greater relative attention is paid to cue C on AC trials than to cue B on AB trials, under typical base-rate designs (e.g. the inverse base-rate effect: Don et al., 2019; and the highlighting effect[2]: Kruschke et al., 2005). Don et al. (2019) measured gaze biases to cues both prior to making a prediction, and during feedback, and assessed how gaze patterns differed based on the global base-rates of the outcomes.

[1] Importantly for the current study, a reduced EXIT model with attentional competition components makes the prediction that the rare predictor C will command greater attention than the common predictor B, whereas as a reduced EXIT model with rapid attention shifts makes the opposite prediction (Paskewitz & Jones, 2020).
[2] In highlighting, AB-O1 trials are learned before the introduction of AC-O2 trials, and the highlighting effect refers to a similar bias in predicting O2 on BC trials at test.

DOI: 10.1177/1747021820985522

Manipulation of the global base-rates should affect the associations between the context and the prevailing common outcomes, where the context comprises the incidental cues related to features of the experimental trials and participating in the experiment more generally.

We will describe the design of Don et al. (2019, see Table 1) in detail since it is highly relevant to the current study. In our *standard* condition, one outcome was always paired with common compounds, while another was always paired with rare compounds, such that overall, the context will be more strongly associated with the common outcome. This was compared to a *balanced* outcome condition, where each outcome was paired with one common compound, and one rare compound, such that each outcome was experienced equally across the course of the experiment, and the context will not be strongly associated with either outcome. This condition has been shown to reduce the strength of the inverse base-rate effect (Don & Livesey, 2017; Don et al., 2019). In the standard condition, we found gaze biases towards C on AC trials both prior to making a prediction, and during feedback. In the balanced condition, there was an equivalent bias to B on AB trials as there was to C on AC trials, prior to making a decision. However, attention during feedback did not differ from that in the standard condition (both conditions showed a bias towards C). These patterns of attention changed differently across training for each stage of the trial. While preferential attention to C (and to B in the balanced condition) prior to making a prediction increased across the course of training, attention to C during feedback was high early in training, and decreased as training progressed.

--- Insert Table 1 about here ---

DOI: 10.1177/1747021820985522

These results highlight two attention-based processes that may contribute to the inverse base-rate effect. In the standard condition, although both B and C are both perfectly predictive cues, B is trained in combination with other cues that also predict the same outcome (A and the context are both more predictive of the common outcome), whereas C is trained with cues that predict a different outcome. As such, C may be considered a more useful predictor than B, as it provides a greater informational advantage over the other cues present. We would expect this predictive advantage for C to grow as learning improves over training. We would also expect a greater informational advantage for B if the context were less predictive of the common outcome (as is the case in the balanced condition). These patterns are borne out in gaze while participants make predictions, and therefore this attention bias may reflect learning cue predictiveness. In addition, there is a large amount of prediction error that occurs on AC trials. As a surprising outcome occurs on these trials, attention may be driven particularly strongly *away* from any discrete cue that generates the prediction error (i.e. the imperfect predictor A) and thus towards the rare predictor C even though the association between C and the rare outcome may still be developing. Thus, attention to C might be enhanced by the larger prediction error that is experienced on rare trials. We would expect this effect, to the extent that it correlates with the magnitude of prediction error, would diminish across training as accuracy improves. This pattern is largely borne out in gaze patterns during the feedback period of the trial. Learned predictiveness and prediction error will of course be linked, as prediction error will decrease as participants learn the predictive relationships across training.

The gaze data reported by Don et al. (2019) demonstrate that overt attention biases have a complex relationship to learning, potentially reflecting several functional properties of

DOI: 10.1177/1747021820985522

competitive learning as the inverse base-rate effect is acquired. However, a question remains as to whether gaze in this instance even reflects attentional changes that are relevant to ongoing selective learning and, if so, which pattern of gaze is more indicative of changes in the selective prioritising of cues during learning. The aim of the current study was to test the relative associability of common and rare predictors. Given the results of Don et al. (2019), we also wanted to test 1) how the relative cue associability changed over the course of training, and 2) whether cue associability differences depend on the use of a context that was more strongly associated with the common outcome. Thus, we tested associability of cues after either short or long training, with training in either standard or balanced conditions. Experiment 1 used a three-stage design similar to that used to assess learned predictiveness effects (e.g., Le Pelley & McLaren, 2003). Following base-rate training, this experiment included a new training phase that paired a previously common predictor (B) with a previously rare predictor (C), followed by a novel outcome in a novel context. If, for example, greater attention is paid to rare predictors than common predictors throughout base-rate training, then rare predictors should be more strongly associated with the novel outcomes than are the common predictors. If cue associability reflects learned attention due to learned predictiveness, we would expect weaker differences in the associability of rare versus common predictors in the balanced condition compared to the standard condition, and stronger biases after long training than short training. If cue associability follows current prediction error, we should expect no differences between standard and balanced conditions, and weaker biases after long training than short training. Our previous demonstrations of the inverse base-rate effect and the effect of using a balanced design have all used relatively long training in which prediction accuracy is high for all trial types by the end of training. However, one of our competing hypotheses about the source of enhanced attention to C

DOI: 10.1177/1747021820985522

assumes that the presence of prediction error (i.e. earlier in training) is important. To establish whether, after *short* training, there is an inverse base-rate effect and whether it is affected by the predictive status of the context, Experiment 2 compared the inverse base-rate effect in standard and balanced conditions after short training.

Experiment 1

Experiment 1 compared the associability of common and rare predictors in standard and balanced conditions after short and long training. The cues that receive greater attention during base-rate training should be learned about more readily in a new learning phase. The design of the experiment is shown in Table 2. In Phase 1, participants were trained with either the standard or balanced design as in Don et al. (2019). Then, in Phase 2, all participants completed a second training phase in which they were presented with new compounds comprising one previously rare predictor and one previously common predictor, paired with a novel outcome, in a novel context. Importantly, in this new learning phase, all compounds were trained in equal base-rates, and each cue was equally predictive of its respective outcome. Thus, any differences in learning about cue-outcome associations in phase 2 would be attributable to changes in their associability as a result of previous base-rate training. It is worth noting that this design does not include the typical test phase to assess the inverse base-rate effect. Kruschke et al. (2005) included test trials before the transfer phase, however this could potentially disrupt the transfer of associability between phases. Instead, learning in Phase 2 was tested using two different kinds of test trial, summation and negation compounds, to provide converging evidence of associability biases (e.g., Livesey et al. 2011). On summation trials (e.g., BE), two cues of the same type in Phase 1 (e.g., previously common predictors) that were paired with the same outcome in Phase 2 were presented together. Thus, the critical comparison is prediction accuracy for the summation

DOI: 10.1177/1747021820985522

compounds composed of previously common predictors compared to the compounds composed of previously rare predictors. On negation trials (e.g., BC), two cues of different types in Phase 1 (e.g., one previously common predictor, one previously rare predictor) that were paired with different outcomes in Phase 2 were presented together. The critical comparison here is the proportion of choice of the outcome that was paired with the previously common predictor compared to choice of the outcome that was paired with the previously rare predictor. Note that these negation trials are the same as the conflicting trials, but here they do not provide a test of the inverse base-rate effect, but assess learning of the contingencies with the new outcomes in Phase 2. If there is a significant attention bias to rare cues in Phase 1 which influences associability in Phase 2, then participants should have greater accuracy on the rare summation trials than common summation trials, and show a greater proportion of choice of the outcome paired with the previously rare predictor in the negation trials.

To determine how associability changes across training, there were two training length conditions, where participants either received 42 repetitions of each common compound, and 14 repetitions of each rare compound in Phase 1 (Experiment 1A), or a shorter training phase with 18 repetitions of each common compound, and six repetitions of each rare compound (Experiment 1B).

Method

Participants

One hundred and ten first-year psychology students at the University of Sydney participated in return for partial course credit. One participant was excluded for not reaching the training criterion during Phase 1. This left 109 participants (70 female, mean age = 19.4, *SD* =

DOI: 10.1177/1747021820985522

3.2) randomly allocated to standard and balanced conditions. Recruitment for the short training conditions (Experiment 1B) occurred after completing recruitment for the long training conditions (Experiment 1A). The final sample included 27 participants in each of the long-standard, long-balanced, and short-standard conditions, and 28 in the short-balanced condition.

Design

The design is shown in Table 2. In Phase 1, participants received either standard or balanced base-rate training. In Phase 2, new compounds comprising one previously common predictor, and one previously rare predictor were paired with one of two novel outcomes, e.g., BI – O3. The test phase assessed learning of the contingencies in Phase 2 using the summation and negation test trials described above.

--- Insert Table 2 about here ---

Apparatus and Stimuli

The experiment was programmed using PsychToolbox for Matlab (Kleiner, Brainard & Pelli, 2007) and was presented using Apple Mac Mini computers attached to 17-inch displays. Experimental stimuli included 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*, with accompanying labels in blue text. Foods were randomly allocated to cues A-L. The four allergic reaction outcomes were randomly allocated from *Headache*, *Nausea*, *Rash* and *Fever*.

DOI: 10.1177/1747021820985522

Procedure

Participants were instructed to assume the role of a doctor whose task was to determine which foods were causing which allergic reactions in their fictitious patients. On each training trial, two cues appeared on the upper half of the screen. After 500ms, the outcome options were presented in boxes on the lower half of the screen, and participants used the mouse to make an outcome prediction. Once an outcome was selected, the selected box turned blue. The outcomes then disappeared and corrective feedback was provided for two seconds. The correct outcome was shown, accompanied by the word “correct” in green, or “incorrect” in red, depending on the accuracy of the prediction.

The position of cues on screen was counterbalanced within each block, and the position of outcomes was counterbalanced across participants. There were three blocks of training with a 3:1 base-rate; each block contained six presentations of each common compound and two presentations of each rare compound (see Table 1). Participants received training in either the standard design, where each outcome was consistently paired with either common or rare compounds, or the balanced design, where each outcome was paired with both a common compound, and a rare compound.

Two versions of the experiment were run consecutively. In Experiment 1A, there were seven blocks of training each containing six repetitions of each common compound, and two repetitions of each rare compound. In Experiment 1B there were three blocks of training, again with six repetitions of each common compound, and two of each rare compound. In Phase 1, participants predicted allergic reaction outcomes for their patient, Mr X. At the beginning of Phase 2, participants were instructed that they would now see a new patient, Miss Y, and were to continue predicting which foods would lead to which allergic reaction. They were informed that

DOI: 10.1177/1747021820985522

Miss Y ate many of the same foods as Mr X, but suffered from different allergic reactions. The Phase 2 compounds contained one previously common predictor, and one previously rare predictor from Phase 1, and each compound was paired with one of two novel outcomes. Trials continued in a similar manner as Phase 1, however each cue compound was presented with equal frequency, and each cue was equally predictive of the outcome with which it was paired. There were three blocks of Phase 2 training in all groups. Each cue compound was presented twice per block, with counterbalanced cue position on the screen. In the test phase, participants were asked to predict which allergic reaction Miss Y was most likely to suffer from, given the presented foods, and to rate their confidence. They were informed they would no longer receive feedback for their responses. On each trial, one of the summation or negation test compounds was presented on the top half of the screen. Participants selected the outcome they thought was most likely by clicking an option, which then turned blue. After selecting an outcome, a linear analogue scale appeared beneath the outcome options, accompanied by the question "How confident are you that this is the correct choice?" Participants rated their confidence on the scale, which ranged from "not at all confident" to "very confident". Responding was self-paced, and participants were able to modify both responses before pressing the space bar to move to the next trial. Each test trial was presented once and in random order. The position of cues on screen was randomised for each trial.

Results

Training Phase 1

Experiment 1A. Training data are presented in Figure 1. For analysis of Phase 1, a 2 x (2) x (7) mixed-measures ANOVA was run with global base-rate group (standard vs. balanced)

DOI: 10.1177/1747021820985522

as the between-subjects factor and trial type (common trials vs. rare trials) and block (1-7) as within-subjects factors. This revealed a main effect of block, $F(6,312) = 147.46$, $p < .001$, $\eta_p^2 = .739$, indicating an improvement in accuracy across training. There was a significant main effect of trial type, $F(1,52) = 95.38$, $p < .001$, $\eta_p^2 = .647$, with greater accuracy for common trials (M = .96, SD = .03) than rare trials (M = .86, SD = .08) overall, and a significant main effect of global base-rate group $F(1,52) = 4.43$, $p = .04$, $\eta_p^2 = .078$, with greater overall accuracy in the standard group (M = 0.92, SD = .04) than the balanced group (M = .90, SD = .05). There was also an interaction between block and trial type, $F(6,312) = 30.28$, $p < .001$, $\eta_p^2 = .368$. Figure 1a suggests that common trials were learned faster than rare trials, but accuracy on the different trial types converged later in training. To further analyse this interaction, we compared the difference in accuracy between common and rare trials, which was greater in the first block of training (*mean difference* = .29, SD = .20) than the final block of training (*mean difference* = .02, SD = .05), $t(53) = 9.10$, $p < .001$, $d = 1.24$.

Experiment 1B. Experiment 1B showed a similar pattern of results to Experiment 1A. There was a main effect of block, $F(2,106) = 166.91$, $p < .001$, $\eta_p^2 = .759$. Accuracy for common trials (M = .91, SD = .07) was higher than accuracy for rare trials (M = .78, SD = .13) overall, $F(1,53) = 77.87$, $p < .001$, $\eta_p^2 = .595$. There was a significant interaction between block and trial type, $F(2,106) = 6.33$, $p = .003$, $\eta_p^2 = .107$. Figure 1b suggests that common trials were again learned faster than the rare trials. The difference in accuracy for common and rare trials was higher in the first block of training (*mean difference* = .19, SD = .24) than the final block of training (*mean difference* =.08, SD = .15), $t(54) = 3.29$, $p = .002$, $d = .443$. There was greater accuracy overall in the standard group (M = .89, SD = .05) than the balanced group (M = .80, SD = .08), $F(1,53) = 23.86$, $p < .001$, $\eta_p^2 = .31$. Additionally, there was a trial type x global base-rate

DOI: 10.1177/1747021820985522

group interaction, $F(1,53) = 24.36$, $p < .001$, $\eta_p^2 = .315$, where the difference in accuracy for common and rare trials was greater in the balanced group (*mean difference* = .20, *SD* = .12) than standard group (*mean difference* = .06, *SD* = .09) Analyses of simple effects indicated this difference was significant in both standard ($t(26) = 3.18$, $p = .004$, $d = 0.61$) and balanced ($t(27) = 8.77$, $p < .001$, $d = 1.66$) groups.

--- Insert Figure 1 about here ---

Training Phase 2

To assess the influence of training length on subsequent learning, accuracy in Phase 2 learning was analysed for Experiment 1A and 1B together. A 2 x 2 x (3) mixed measures ANOVA was run with training length and group as between-subject factors, and block as a within-subjects factor, which showed a main effect of block, $F(2,210) = 130.33$, $p < .001$, $\eta_p^2 = .554$. There was a significant main effect of global base-rate group, $F(1,105) = 4.01$, $p = .048$, $\eta_p^2 = .037$, with overall accuracy greater in the standard group ($M = .78$, $SD = .11$) than the balanced group ($M = .73$, $SD = .15$). There was also an interaction between block and training length, $F(2,210) = 3.11$, $p = .047$, $\eta_p^2 = .029$. Figure 1c suggests a difference in the rate of learning. However, this interaction is difficult to interpret as there were no significant differences between training length groups in any block of Phase 2 training, highest $t(107) = 1.73$, $p = .086$, $d = 0.33$.

Test

DOI: 10.1177/1747021820985522

For analyses of the test phase we include both frequentist and Bayesian tests, which can be interpreted as the odds in favour of the alternative hypothesis (Wagenmakers et al., 2018). The Bayesian tests were run in JASP using Bayesian ANOVAs or t-tests with default priors. Bayes factors for the main effects indicate the likelihood of the data given the main effects model relative to a null model (BF_{10}). Bayes Factors on interaction effects indicate evidence for the interaction by comparing models including the interaction effect with models excluding the effect (BF_{incl}; Rouder et al., 2017).

Summation trials. Accuracy on summation trials is shown in Figure 2A. A 2 x 2 x (2) ANOVA with training length (long vs. short) and global base-rate group (standard vs. balanced) as a between subjects factors, and cue type (previously common vs. rare predictors) as a within subjects factor revealed a significant main effect of cue type, $F(1,105) = 11.43$, $p = .001$, $\eta_p^2 = .098$, $BF_{10} = 45.65$, such that overall, participants were more accurate for cue compounds comprising previously rare predictors than previously common predictors. There was no main effect of training length, $F < 1$. However, there was a significant interaction between cue type and training length, $F(1,105) = 4.65$, $p = .033$, $\eta_p^2 = .042$, $BF_{incl} = 2.0$, indicating that this benefit for rare over common predictors was stronger after three blocks of training (*mean difference* = .21, *SD* = .41) than after seven blocks (*mean difference* = .05, *SD* = .38). To further analyse this interaction, two separate ANOVAs for each training length group showed a significant effect of cue type after short training, $F(1,53) = 14.45$, $p < .001$, $\eta_p^2 = .214$, $BF_{10} = 384.68$, but not after long training, $F < 1$, $BF_{10} = 0.29$. Interestingly, there was no significant main effect of global base-rate group, $F(1,105) = 1.03$, $p = .312$, $\eta_p^2 = .01$, $BF_{10} = 0.26$, and no interaction between cue and global base-rate group $F < 1$, $BF_{incl} = 0.26$, nor were there any significant main effects or interactions with global base-rate group in either training length condition when analysed

DOI: 10.1177/1747021820985522

separately, highest $F(1,53) = 1.13$, $p = .292$, $\eta_p^2 = .021$, $BF_{10} = 0.32$. These results indicate that the effects of cue type were not significantly stronger in the standard group than the balanced group.

Negation. The proportion of choice of each outcome is shown in Figure 2B. As these proportions are complementary, analyses focused on the proportion of choice of the outcome paired with the previously rare predictor. Overall, participants' choices were significantly biased towards the outcome paired with the previously rare predictor (i.e. the proportion of choice of the outcome paired with the previously rare predictor was greater than .5), $t(108) = 4.18$, $p < .001$, $d = 0.40$, $BF_{10} = 285.97$. In a 2 x 2 between-subjects ANOVA, there were no significant main effects of training length, $F(1,105) = 2.39$, $p = .125$, $\eta_p^2 = .022$, $BF_{10} = 0.58$, or global base-rate group, $F(1,105) = 1.63$, $p = .205$, $\eta_p^2 = .015$, $BF_{10} = 0.41$, or interaction between training length and global base-rate group, $F < 1$, $BF_{incl} = 0.27$. Although there was no significant effect of training length condition, based on the significant interaction in the summation results, the effect was analysed separately for each group. There were no significant global base-rate group differences in either training length condition, Fs < 1, BFs < 0.39. There was a significant bias towards the rare predictor in the short training group, $t(54) = 4.05$, $p < .001$, $d = 0.55$, $BF_{10} = 141.09$, but this bias did not reach significance in the long training group, $t(53) = 1.88$, $p = .066$, $d = .25$, $BF_{10} = 0.75$. Confidence ratings on summation and negation trials are shown in Table 3.

--- Insert Figure 2 about here ---

--- Insert Table 3 about here ---

Discussion

DOI: 10.1177/1747021820985522

Experiment 1 demonstrates greater associability for rare predictors than common predictors overall. Accuracy was better on summation test trials comprising previously rare predictors compared to those comprising previously common predictors. On negation trials, there was a significant bias in choice favouring the outcome paired with the previously rare predictor over the outcome paired with the previously common predictor. These findings suggest that participants pay greater attention to rare predictors during training, which facilitates subsequent learning about those cues.

Although the bias for rare predictors appeared to be weaker in the balanced group than the standard group in several conditions, this difference did not reach significance, and Bayes factors provided more evidence for the null hypothesis. There was an effect of training length on associability effects as measured by the summation test trials, where the advantage for previously rare predictors was greater following short training than following long training. Although this effect was not significant in the negation trials, outcome choice followed a similar numerical pattern, and neither group showed a significant effect after long training.

We will reserve further theoretical interpretation of these results for the General discussion. For now, we note that the presence of particularly strong associability biases after short training (evident regardless of the predictive status of the context) warrants a test for the presence of choice biases after short training. This was therefore the aim of Experiment 2.

Experiment 2

We have reliably observed an inverse base-rate effect, as well as a difference in the strength of the effect between standard and balanced groups, after longer training used in Experiment 1 (Don & Livesey, 2017; Don et al., 2019). While the inverse base-rate effect has been

DOI: 10.1177/1747021820985522

demonstrated after various amounts of training across studies, we have not determined whether the inverse base-rate effect, or the difference in the effect as a result of global outcome base-rate differences, occurs after a short amount of training in this particular design and procedure. In Experiment 1, we observed an associability bias for rare predictors after short training, and this effect did not differ between groups. Thus, the aim of Experiment 2 was to compare the inverse base-rate effect in standard and balanced conditions after the same relatively short training phase used in Experiment 1. This will allow us to determine whether associability effects relate to the strength of the inverse base-rate effect, and whether associability effects in training precede the emergence of the effect.

Method

Participants

Forty-nine undergraduate students from the University of Sydney participated in return for partial course credit (29 female, mean age = 23.6, *SD* = 7.0), and were randomly allocated to standard (n = 24) and balanced (n = 25) groups.

Apparatus & Stimuli

Apparatus and stimuli were identical to those used in Experiment 1.

Procedure

The training phase was identical to Phase 1 in the short training condition of Experiment 1. The test phase followed immediately after training, and proceeded in a manner similar to Experiment 1, but using the test trials shown in Table 1. Participants were instructed to use the knowledge that they had gained so far to respond to trials without feedback. On each trial, one,

DOI: 10.1177/1747021820985522

two, or three cues appeared on the upper half of the screen, and participants selected the outcome they thought was most likely, and rated their confidence. Participants were able to modify their responses before proceeding to the next trial.

Results

Training

Response accuracy during training is shown in Figure 3. A 2 x 2 x (3) mixed measures ANOVA was run with global base-rate group (standard vs. balanced) as the between-subjects factor and trial type (common vs. rare) and block (1-3) as within-subjects factors. This revealed a significant main effect of block, indicating an increase in accuracy across training, $F(2,94) = 110.10$, $p < .001$, $\eta_p^2 = .701$. There was a significant main effect of trial type, indicating greater accuracy on common trials ($M = .91$, $SD = .06$) than rare trials ($M = .73$, $SD = .14$), $F(1,47) = 130.16$, $p < .001$, $\eta_p^2 = .735$. A significant interaction between block and trial type suggests common trials were learned faster than rare trials, $F(2,94) = 11.17$, $p < .001$, $\eta_p^2 = .192$. The difference in accuracy for common and rare trials was greater in the first block ($M = .27$, $SD = .22$) than the final block ($M = .09$, $SD = .17$) of training, $t(48) = 4.74$, $p < .001$, $d = 0.68$. A significant interaction between trial type and global base-rate group also suggests that the difference in accuracy for common and rare trials was greater in the balanced group (*mean difference* = .23, $SD = .11$) than the standard group (*mean difference* = .12, $SD = .11$), $F(1,47) = 12.32$, $p = .001$, $\eta_p^2 = .208$. Further analysis of simple effects indicated that the difference between overall common and rare trial accuracy was significant in both the standard group ($t(23) = 5.59$, $p < .001$, $d = 1.14$), and balanced group ($t(24) = 10.56$, $p < .001$, $d = 2.11$). In addition, the difference in accuracy for common and rare trials remained significant in the final block of

DOI: 10.1177/1747021820985522

training in the balanced group ($t(24) = 3.49$, $p = .002$, $d = 0.70$), but not the standard group ($t(23) = 1.91$, $p = .07$, $d = 0.39$).

--- Insert Figure 3 about here ---

Test

The proportion of rare outcome choice on each of the critical trial types is shown in Figure 4. Analyses focused on these trials, but the proportion of rare outcome choice and mean confidence ratings for each test trial type is shown in Table 5. The proportion of rare outcome choices for each trial type was compared against a chance level of 0.5 using a one-sample t-test. An inverse base-rate effect is present if rare outcome choices are significantly above chance. There was a significant inverse base-rate effect in the standard group, $t(23) = 5.06$, $p < .001$, $d = 1.03$, $BF_{10} = 615.99$, but not in the balanced group, $t(24) = 0.89$, $p = .38$, $d = 0.18$, $BF_{10} = 0.30$. An independent samples t-test indicated that this group difference was significant, $t(47) = 4.26$, $p < .001$, $d = 1.22$, $BF_{10} = 222.67$. On imperfect trials, choices were significantly common-biased in both groups, lowest $t(24) = 4.09$, $p < .001$, $d = 0.82$, $BF_{10} = 74.0$, and there was no significant group difference, $t(47) = 1.7$, $p = .096$, $d = 0.49$, $BF_{10} = 0.92$. On combined trials, choice did not differ from chance in the standard group, $t(23) = 0.81$, $p = .426$, d = 0.17, $BF_{10} = 0.29$, but was significantly common biased in the balanced group, $t(24) = 4.94$, $p < .001$, d = 0.99, $BF_{10} = 514.68$. The group difference was significant, $t(47) = 3.13$, $p = .003$, $d = .89$, $BF_{10} = 12.59$.

--- Insert Figure 4 about here ---

--- Insert Table 4 about here ---

DOI: 10.1177/1747021820985522

Discussion

On the critical conflicting (BC) trials, the difference in choice between standard and balanced conditions after short training appear to be as pronounced, if not more so, than what we have previously observed after longer training (*d* = 1.22 in the current study compared to *d* = 0.55 in Don & Livesey, 2017, and *d* = 0.62 in Don et al., 2019). The inverse base-rate effect was relatively strong in the standard group (*d* = 1.03, compared to *d* = 0.89 in Don & Livesey, 2017, and *d* = 0.78 in Don & Livesey, 2019). Choice was numerically biased towards the common outcome in the balanced group, although this did not significantly differ from chance. This differs from previous studies where we have typically observed a small rare bias in the balanced group (*d* = 0.20 in Don & Livesey, 2017, and *d* = 0.22 in Don et al., 2019). This result may be due to the difference in accuracy for AC trials by the end of training – accuracy was near asymptote in the standard group, but not the balanced group. In any case, the clear effect in the standard group suggests that the strong cue associability effect observed in Experiment 1 does not precede the inverse-base rate effect, and the difference between groups suggests that the associability differences observed in Experiment 1 after short training are not fully sufficient to produce an inverse base-rate effect, since there is no such effect in the balanced condition despite there being evidence of associability biases in this condition in Experiment 1.

General Discussion

Experiment 1 showed better learning about cues that were previously rare predictors than previously common predictors in a new learning phase with novel outcomes. This change in associability indicates that greater attention was paid to rare predictors than common predictors during the first phase of training, and is therefore consistent in this respect with the attentional account of the inverse base-rate effect offered by Kruschke (1996; 2001a), and previous evidence

DOI: 10.1177/1747021820985522

of greater attention to rare predictors (Don et al., 2019; Wills, Lavric, Hemmings & Surrey, 2014). This result complements a similar finding in the highlighting effect, in which AB-O1 trials are trained prior to the introduction of AC-O2 trials, and a similar preference for O2 on BC trials is observed at test. Kruschke (2005) found a negative transfer effect, where learning was poorer for new predictive cues when they were paired with previously late predictors (C) than previously early predictors (B), suggesting continued attention to C in new learning. The current study demonstrates the associability of C is increased when AB and AC trials are trained concurrently.

The length of training appeared to have some effect on cue associability; a substantial associability bias towards C over B was present on both summation and negation tests after short training, whereas neither test trial yielded strong evidence for this effect after longer training, and a significant effect of training length was evident for summation tests trials. This is consistent with the idea that there is an attention advantage for C while AC trials are associated with relatively high prediction error early in training. In the EXIT model, for instance, on experiencing prediction error, attention is quickly shifted towards the cue that will minimize that error and away from the predictive cues that contribute to it. It should also be noted that although these results suggest little benefit for rare predictors after seven blocks of training, the inverse base-rate effect is reliably demonstrated following training of this length or greater. This general finding is not necessarily incompatible with the results of the current experiment because attention biases early in training may be sufficient to develop stronger learning for the rare predictor, which might be maintained throughout extended training even if the attentional bias itself is not. Studies that have reversed or altered the base-rates of contingencies throughout training tend to show a preference for the early rare over the early common outcome on

DOI: 10.1177/1747021820985522

conflicting trials, indicating the importance of early relative frequencies for the effect (Medin & Bettger, 1991; Kruschke, 2009).

We assessed the effect of context associations on attention transfer by comparing the standard condition, in which the context will be strongly associated with the common outcome, to a balanced condition, where the context will not be strongly associated with either outcome. Our previous work (Don & Livesey, 2017; Don et al., 2019) has indicated that this manipulation has a strong effect on the magnitude of the inverse base-rate effect. Although the transfer effect in Experiment 1 appeared numerically weaker in the balanced group on some tests, group differences between standard and balanced conditions were not significant either overall or for individual training lengths, on either summation or negation tests. Experiment 2 demonstrated a robust inverse base-rate effect after short training in the standard condition, suggesting that this attentional bias after short training should not necessarily be considered a precursor of the rare choice bias, but possibly something that emerges with it. In addition, the inverse base-rate effect was strongly affected by balancing the global frequency of the outcomes, such that choice proportion favoured neither rare nor common outcomes, but the attention bias was still present at short training in the balanced condition and was not affected by the standard vs balanced manipulation. This suggests that the associability bias alone is not sufficient to produce the inverse base-rate effect, though it is possibly one of its necessary conditions. This may add to a list of conditions that appear to be necessary but not sufficient for the effect to occur, including prediction error during training (Kruschke, 2001a; Medin & Edelson, 1988; Wills et al., 2014), and the presence of global outcome base-rate differences (Don & Livesey, 2017). Granted, we did not assess the magnitude of the inverse base-rate effect and cue-associability within the same experiment and therefore cannot directly assess associations between the two on a participant

DOI: 10.1177/1747021820985522

level. This was done to avoid potential interference between separate test phases, however future research could measure both within the same experiment with counterbalanced test orders.

Given past evidence of associability in the learned predictiveness effect, it would be fair to expect that the kind of attention that influences future learning would be best reflected by long term learned attention to cues. Yet the change in transfer effects across training does not reflect this. Transfer effects instead appear to follow a similar pattern to eye gaze during feedback seen in Don et al. (2019). That is, attention biases to C were stronger earlier in training than later in training, and there was little difference between standard and balanced conditions. We have speculated that this pattern of attention reflects the current state of prediction error, rather than learned attention based on predictiveness. This would leave the current results seemingly at odds with a wealth of literature on associability and attention in the learned predictiveness effect. In the learned predictiveness effect, Phase 1 training usually proceeds to a point where participant predictions are highly accurate (i.e. there is very little prediction error, at least in the participants' overt predictions), and there are highly replicable transfer effects where previously predictive cues are learned about more readily than previously non-predictive cues (Le Pelley & McLaren, 2003; Le Pelley, Turnbull, Reimers & Knipe, 2010; Don & Livesey, 2015; Shone, et al., 2015). These effects are also associated with changes in pre-decision gaze biases (e.g., Le Pelley et al., 2011). Thus, in this literature, there is a strong link between transferred attention and learned predictiveness (and *not* current prediction error). However, the notion that attention might reflect current prediction error is consistent with recent findings that suggest uncertainty about the outcome is associated with sustained attention to cues (Beesley, Nguyen, Pearson & Le Pelley, 2015).

DOI: 10.1177/1747021820985522

The most obvious reason why our results may differ is because we pitted two perfect predictors against each other. Arguably, this might nullify any associability transfer effects attributable to learned predictiveness. Despite the large body of research on the effect, the precise operations of the learned predictiveness effect are still not well known. Some results suggest, for instance, that competition among cues with different predictive validity (i.e. *relative* predictiveness) is completely unnecessary for the effect (Kattner, 2015; Le Pelley et al., 2010; Livesey et al., 2011), suggesting that the absolute predictiveness of each cue determines their attention in new learning. If this were true then it would be reasonable to assume that B and C cues receive the same benefits from learned predictiveness effects in new learning and any bias towards C is attributable to other differences, such as those driven by its relative utility in resolving prediction error on the most recent trials. Although we have not focused on the comparison of formal models here, we have previously shown that Mackintosh's (1975) model predicts greater attention to B than C, and that overt attention does not follow this pattern. However, in developing his model, Mackintosh (1975) outlined formal assumptions about cue processing changes as a consequence of learning (cue *associability*, specifically) but remained agnostic about how these changes will manifest in patterns of overt attention or orienting, which he noted were outside the scope of his formal analysis. Here we confirm the same general pattern of prioritised attention to C for associability. Thus, this is again more consistent with the predictions of the EXIT model than the Mackintosh model. The results will inform any discussion of which theoretical mechanisms are necessary in more complex mechanisms like EXIT (e.g. see Paskewitz & Jones, 2020).

Notwithstanding the hypothesised processes discussed above, it must be acknowledged that the common and rare cues (by their very nature) differ in their frequency of exposure and

DOI: 10.1177/1747021820985522

this by itself can have an effect on cue salience. Cues that are presented without any consequences appear to lose associability when they are later paired with a meaningful outcome (the latent inhibition effect, see Holmes & Harris, 2010 for a review). It is possible that this process still occurs even when cues are presented with reliable consequences (e.g. Jones & Haselgrove, 2013; Kaye & Pearce, 1984). Latent inhibition effects have been notoriously difficult to demonstrate in humans (see Byrom et al., 2018), but if they were to have a substantial impact on cue associability in this type of explicit learning task then they could contribute to greater attention being paid to the rarer of two predictive cues. This would not be sufficient to explain the inverse base-rate effect itself (it does not explain the difference between standard and balanced conditions, for instance) but it might be sufficient to explain why one would attend to C more than B, that is, on the basis of relative novelty alone. Future research may be needed to tease apart contributions of prediction error and mere novelty on this associability effect.

In summary, we have demonstrated an attention bias to rare predictors that persists into new learning. This bias was stronger following short training than following longer training, and was unaffected by differences in global outcome base-rates. However, global base-rates have a clear effect on choice biases that constitute the inverse base-rate effect, even after short training. Thus, it appears the kind of attention bias we have observed here is not sufficient for producing the inverse base-rate effect. In addition, the pattern of associability effects closely matched the pattern of eye gaze observed during feedback in Don et al. (2019), and may be a reflection of current prediction error. While prior research has shown relationships between associability and overt attention prior to making a decision, the current results suggest the relationship between associability and attention is still not well understood, and will require further research in order to make meaningful progress towards theory development.

DOI: 10.1177/1747021820985522

Author note

The study was funded by an Australian Postgraduate Award granted to HJD. Summarised data is available on the Open Science Framework: https://osf.io/4362u/

DOI: 10.1177/1747021820985522

References

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

Byrom, N. C., Msetfi, R. M., & Murphy, R. A. (2018). Human latent inhibition: Problems with the stimulus exposure effect. *Psychonomic Bulletin & Review, 25*, 2102-2118.

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

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.

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 and Cognition, 45*, 143.

Easdale, L. C., Le Pelley, M. E., & Beesley, T. (2019). The onset of uncertainty facilitates the learning of new associations by increasing attention to cues. *Quarterly Journal of Experimental Psychology, 72*, 193-208.

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

DOI: 10.1177/1747021820985522

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.

Jones, P. M., & Haselgrove, M. (2013). Overshadowing and associability change: Examining the contribution of differential stimulus exposure. *Learning & behavior, 41*, 107-117.

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.

Kaye, H., & Pearce, J. M. (1984). The strength of the orienting response during Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 10*, 90-109.

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.

DOI: 10.1177/1747021820985522

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. (2009). *Highlighting: A canonical experiment*. In B. H. Ross (Ed.), The psychology of learning and motivation (Vol. 51, pp.153–185).

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.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2011). Overt attention and predictiveness in human contingency learning. *Journal of Experimental Psychology: Animal Behaviour Processes, 37*, 220–9.

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

DOI: 10.1177/1747021820985522

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.

Livesey, E. J., & McLaren, I. P. L. (2007). Elemental associability changes in human discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes, 33*, 148.

Livesey, E. J., Don, H. J., Uengoer, M., & Thorwart, A. (2019). Transfer of associability and relational structure in human associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*, 125.

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).

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.

DOI: 10.1177/1747021820985522

Mitchell, C. J., Griffiths, O., Seetoo, J., & Lovibond, P. F. (2012). Attentional mechanisms in learned predictiveness. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 191.

O'Bryan, S. R., Worthy, D. A., 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.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E. J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22*, 304.

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

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.

Wagenmakers, E.J. et al. (2018). Bayesian inference for psychology Part II: Example applications with JASP. *Psychonomic Bulletin and Review, 25*, 58-76.

DOI: 10.1177/1747021820985522

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.

DOI: 10.1177/1747021820985522

Figure Captions

Figure 1. Accuracy in training phase 1 of a) Experiment 1A and b) Experiment 1B, and c) phase 2 for all participants. Error bars indicate standard error of the mean.

Figure 2. A) Accuracy in recalling the correct outcome paired with previously common and previously rare predictors in each group. B) Choice of the outcome paired with the previously common or previously rare predictor in each group (note that common and rare choice proportions on negation trials are complementary and thus sum to 1). Error bars indicate standard error of the mean.

Figure 3. Response accuracy during training for each trial type in the standard and balanced groups.

Figure 4. Proportion of rare choice on imperfect, conflicting and combined test trials in standard and balanced groups following three blocks of base-rate training.

DOI: 10.1177/1747021820985522

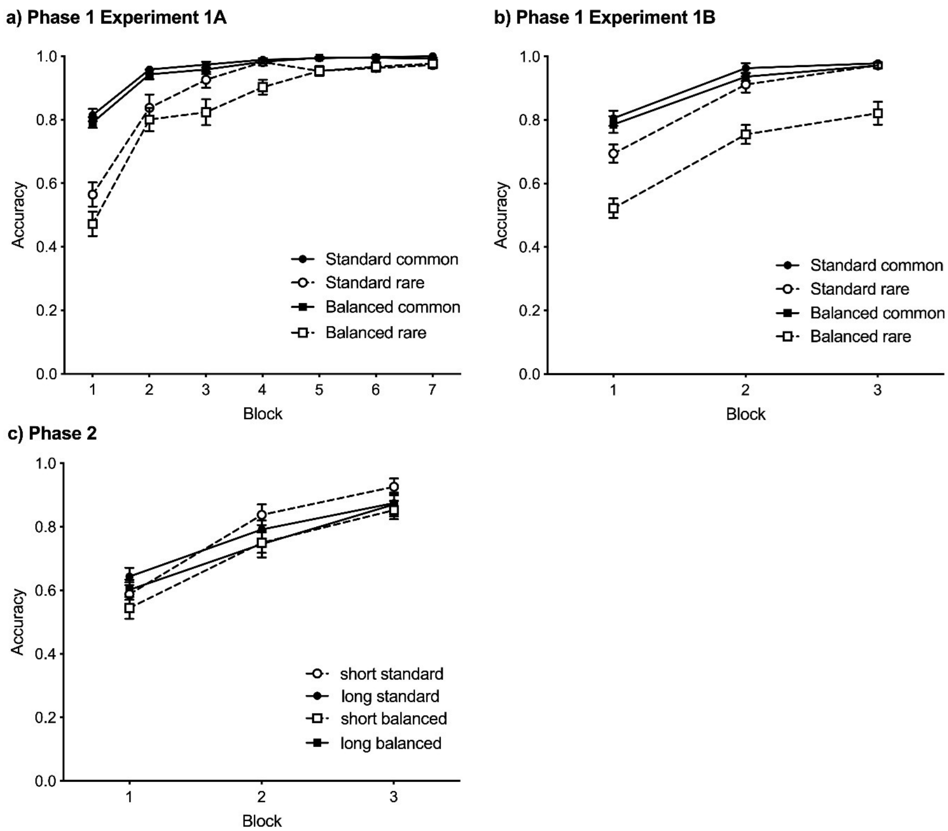

Figure 1. Accuracy in training phase 1 of a) Experiment 1A and b) Experiment 1B, and c) phase 2 for all participants. Error bars indicate standard error of the mean.

164x139mm (220 x 220 DPI)

DOI: 10.1177/1747021820985522

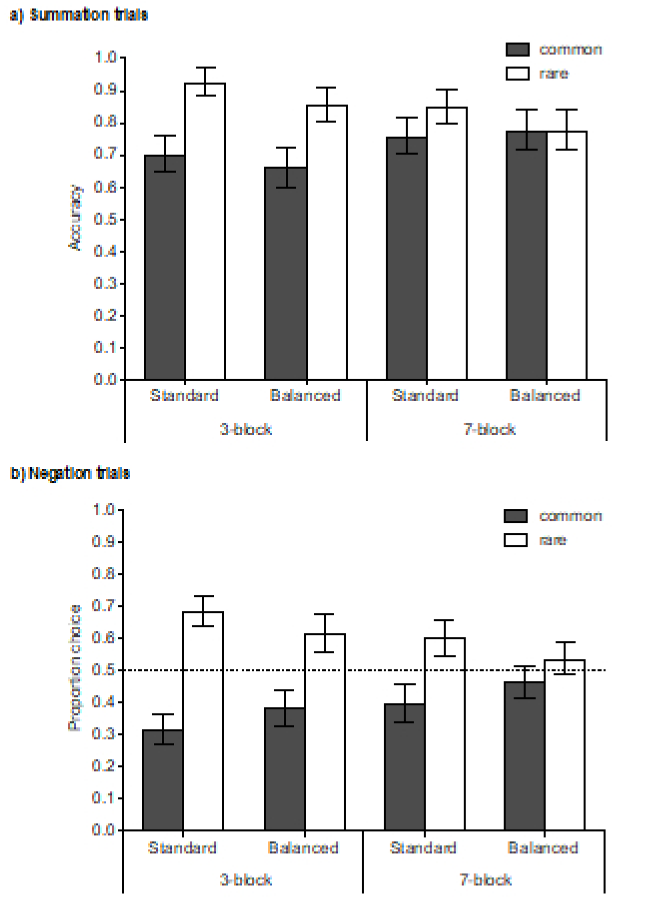

Figure 2. A) Accuracy in recalling the correct outcome paired with previously common and previously rare predictors in each group. B) Choice of the outcome paired with the previously common or previously rare predictor in each group (note that common and rare choice proportions on negation trials are complementary and thus sum to 1). Error bars indicate standard error of the mean.

DOI: 10.1177/1747021820985522

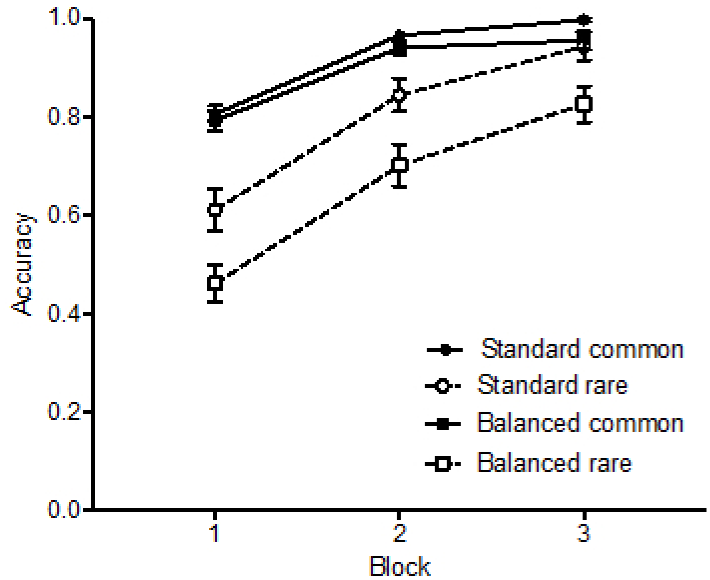

Figure 3. Response accuracy during training for each trial type in the standard and balanced groups.

DOI: 10.1177/1747021820985522

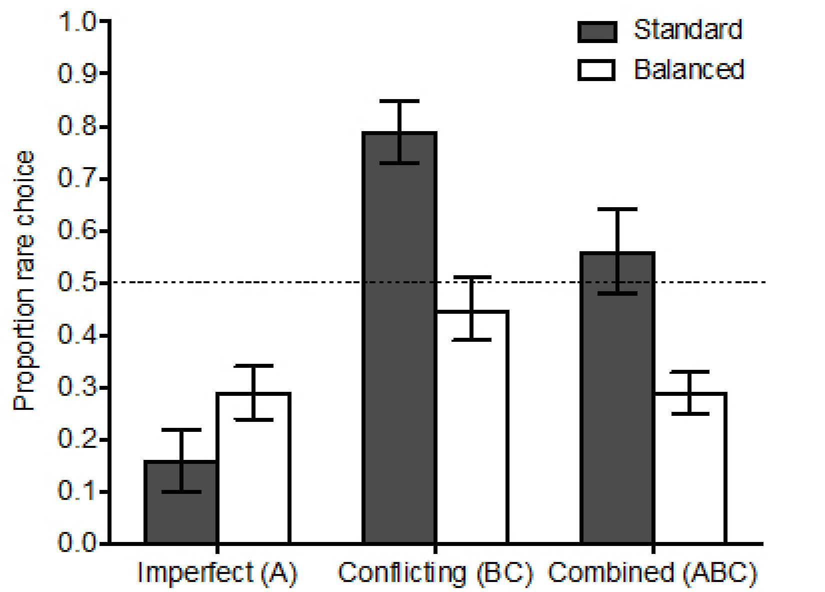

Figure 4. Proportion of rare choice on imperfect, conflicting and combined test trials in standard and balanced groups following three blocks of base-rate training.

DOI: 10.1177/1747021820985522

Table 1.

Design of Don, Beesley & Livesey (2019) and Experiment 2

	TRAINING						
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

DOI: 10.1177/1747021820985522

Table 2.

Experimental design used in Experiment 1

TRAINING PHASE 1					
Group	Base-rate	Trials			
Standard	3	AB – O1	DE – O1	GH – O1	JK – O1
	1	AC – O2	DF – O2	GI – O2	JL – O2
Balanced	3	AB – O1	DE – O2	GH – O1	JK – O2
	1	AC – O2	DF – O1	GI – O2	JL – O1
TRAINING PHASE 2					
	1	BI - O3	CH - O4	EL - O3	FK - O4
TEST PHASE					
Trial type		Trials			
Summation		BE	HK	CF	IL
Negation		BC	EF	HI	LK

Note: Letters refer to individual food cues, O1-O4 refer to different allergic reaction outcomes.

DOI: 10.1177/1747021820985522

Table 3

Confidence ratings for summation and negation trials in Experiment 1.

Group		Summation trials		Negation trials
		Common	Rare	
Standard	*3-block*	66.16	77.73	67.62
	7-block	44.58	56.39	60.49
Balanced	*3-block*	54.27	68.41	62.13
	7-block	39.61	40.49	51.26

DOI: 10.1177/1747021820985522

Table 4

Proportion of rare outcome choices and confidence ratings for each test trial in Experiment 2.

	Standard		**Balanced**	
Test Trial	Proportion choice	Confidence	Proportion choice	Confidence
Imperfect	0.16	68.91	0.29	60.56
Conflicting	0.79	68.84	0.45	59.78
Combined	0.56	66.45	0.29	62.72
Common predictor	0.02	78.82	0.08	68.63
Rare predictor	0.93	78.71	0.80	67.87
Common compound	0.01	94.27	0.03	91.00
Rare compound	0.96	89.15	0.88	79.21

DOI: 10.1177/1747021820985522