Psychonomic Bulletin & Review (2021) 28:1142–1163
https://doi.org/10.3758/s13423-020-01870-0

THEORETICAL REVIEW

Hearing hooves, thinking zebras: A review of the inverse base-rate effect

Hilary J. Don[1,2,3] · Darrell A. Worthy[3] · Evan J. Livesey[1]

Accepted: 20 December 2020 / Published online: 10 February 2021

Abstract

People often fail to use base-rate information appropriately in decision-making. This is evident in the *inverse base-rate effect*, a phenomenon in which people tend to predict a rare outcome for a new and ambiguous combination of cues. While the effect was first reported in 1988, it has recently seen a renewed interest from researchers concerned with learning, attention and decision-making. However, some researchers have raised concerns that the effect arises in specific circumstances and is unlikely to provide insight into general learning and decision-making processes. In this review, we critically evaluate the evidence for and against the main explanations that have been proposed to explain the effect, and identify where this evidence is currently weak. We argue that concerns about the effect are not well supported by the data. Instead, the evidence supports the conclusion that the effect is a result of general mechanisms that provides a useful opportunity to understand the processes involved in learning and decision making. We discuss gaps in our knowledge and some promising avenues for future research, including the relevance of the effect to models of attentional change in learning, an area where the phenomenon promises to contribute new insights.

Keywords Inverse base-rate effect · Human associative learning · Attention in learning · Decision making

Introduction

Many of our daily decisions involve at least some guesswork – take any situation in which we may want to use environmental cues to predict an outcome, whether it be deciding whether to take an umbrella to work based on the current weather, choosing between unfamiliar restaurants based on their menus, or diagnosing an illness based on the patient's symptoms. Most cues we encounter are not fully predictive of the outcomes with which they are associated, and most situations to which we want to generalize our knowledge of the world are not exactly the same as what we have experienced in the past. Consequently, people are often required to make decisions based on ambiguous information. For instance, doctors make diagnoses based on symptoms that are associated with several different conditions. In these cases, the base-rates, or relative frequencies, of events provide an important source of information when making decisions. Indeed, a widely known aphorism in medical circles states "when you hear hoofbeats, think of horses, not zebras", which serves as a reminder that common diagnoses are more probable than rare ones. Yet, several studies have suggested that people often fail to show adequate sensitivity to base-rates.

Base-rate neglect is a phenomenon in which base-rate information is underweighted in favour of more specific information about the individual case (Bar-Hillel, 1980; Bar-Hillel & Fischhoff, 1981; Kahneman & Tversky, 1973). There are of course situations in which favouring specific local information and neglecting base-rates is beneficial for judgements. However, when local information is ambiguous or uninformative, neglecting base-rates may lead to irrational decision-making. For example, in the classic *lawyer-engineer problem*, groups of participants were presented with written personality descriptions and were asked to judge the likelihood that the person described was a lawyer or an engineer (Kahneman &

✉ Hilary J. Don
h.don@ucl.ac.uk

[1] School of Psychology, The University of Sydney, Sydney, NSW, Australia

[2] Division of Psychology and Language Sciences, University College London, London, UK

[3] Psychological and Brain Sciences, Texas A&M University, College Station, TX, USA

Tversky, 1973). Participants were either instructed that the description was randomly drawn from a population of 70 lawyers and 30 engineers, or from a population of 30 lawyers and 70 engineers. Participants' judgements of whether each description referred to a lawyer or engineer were more likely to be based on the stereotypical personality characteristics described than the base-rates of each profession within the sample. Even when given ambiguous, uninformative descriptions, when base-rates should be most informative, participants underweighted this information, predicting a 50% probability of either profession.

These examples of base-rate neglect tend to be observed in explicit decision-making tasks, where base-rate information is provided in a summary statistic. Yet, insensitivity to base-rates appears to be dependent on task conditions (see Koehler, 1996, for a review). Several studies have shown that decisions are more likely to be consistent with base-rates when they are acquired through trial-by-trial experience (Butt, 1988; Christensen-Szalanski & Beach, 1982; Christensen-Szalanski & Bushyhead, 1981; Manis, Dovalina, Avis, & Cardoze, 1980). This kind of direct experience is assumed to make the base-rates more salient, and therefore more likely to be used. Medin and colleagues suggest that the use of base-rate summaries in these text-based tasks fails to consider the influence of learning processes that may allow base-rate information to be incorporated implicitly (Medin & Bettger, 1991; Medin & Edelson, 1988). The *inverse base-rate effect* is therefore particularly noteworthy, as it not only demonstrates a choice bias that goes *against* the underlying base-rates when faced with ambiguous information, but also demonstrates a failure to use base-rates despite acquisition through experience (Medin & Edelson, 1988).

The inverse base-rate effect

To illustrate the inverse base-rate effect, imagine a doctor learning to diagnose diseases on the basis of exhibited symptoms. Over time, they learn that all patients with the symptoms headache and nausea have the *common* disease "midosis", and all patients with the symptoms headache and fever have the *rare* disease "coralgia". A new patient then presents with nausea and fever. Which disease should be diagnosed? Here, both nausea and fever are equally predictive of their respective diseases, and therefore the specific symptoms do not provide evidence in favour of one disease over the other. However, midosis occurs much more frequently than coralgia, and thus a rational response considering the base-rates would be to predict midosis. Yet, given this combination of conflicting cues, most people tend to predict the rare disease. This choice demonstrates a preference for the *less* frequent outcome, and is therefore termed the *inverse base-rate effect*. This effect was first reported by Medin and Edelson (Medin & Edelson, 1988; see also Binder & Estes, 1966) in a contingency learning task where participants played the hypothetical role of a doctor, like the scenario described above. In their task, participants learned symptom-disease contingencies on a trial-by-trial basis. All patients with symptom A and symptom B had disease 1 (AB-O1), while all patients with symptom A and symptom C had disease 2 (AC-O2). Instances of O1 occurred three times as often as instances of O2. Symptom A is therefore an *imperfect* predictor, as it is paired with both diseases. Symptom B is a perfect predictor of the common disease, O1 (hereafter the *common predictor*), and symptom C is a perfect predictor of the *rare* disease, O2 (hereafter the *rare predictor*).[1] After learning these contingencies, participants completed a transfer phase including several new combinations of the trained symptom cues. Most critically, when participants are presented with a pair of *conflicting* symptoms, BC, they tend to diagnose the rare disease. It is these conflicting trial types that yield the inverse base-rate effect. However, almost all studies of the effect also report responses for a series of other transfer trials, which, taken as a whole, reveal the particular circumstances in which the inverse base-rate occurs. As these trials demonstrate the ways in which people make decisions when presented with ambiguous information, we will first define each of the most commonly used types of transfer trials, and summarise the patterns of predictions that they typically elicit. These trial types comprise the key elements of the basic design shown in Table 1, where letters represent individual cues and O1, O2, etc. represent individual outcomes. In the original Medin and Edelson (1988) task, there were three repetitions of this design, with different cues and outcomes for each repetition.

The conflicting $(BC)^2$ compound trials are composed of a perfect common outcome predictor (B) and a perfect rare outcome predictor (C) that shared a cue during training (A). The inverse base-rate effect is indicated by greater choice of the associated rare outcome on these trials. Several studies have also included between-compound conflicting cues, which pair a common and rare predictor that did not share a cue during training. These trials also typically lead to a preference for the rare outcome, although the preference is sometimes slightly numerically weaker (Bohil, Markman, & Maddox, 2005; Kalish, 2001; Kruschke, 1996; Lamberts & Kent, 2007).

The imperfect cue A is associated with both the common and the rare outcome, but typically elicits greater common outcome responses, consistent with base-rate use. These trials are useful to assess normative use of the base-rates of the relevant outcomes, and from a mechanistic perspective, the

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

[2] The descriptive (e.g., conflicting) and abstract (e.g., BC) labels for these transfer trials are used interchangeably throughout this review.

Table 1 Basic inverse base-rate task design

Training phase		Transfer phase	
Base-rate	Trials	Type	Trials
3	A**B** – O1	Imperfect Predictors	A?
1	A<u>C</u> – O2	Conflicting transfer	**B**<u>C</u>?
		Combined transfer	A**B**<u>C</u>?

Note: A – C represent different symptom cues, O1 – O2 represent different disease outcomes. AB – O1, for example, indicates that symptom A and symptom B predicted disease O1

Cues in bold indicate perfect predictors of common outcomes, underlined cues indicate perfect predictors of rare outcomes

The base-rate column refers to the relative number of presentations of each trial type during training, such that AB – O1 occurs three times as often as AC – O2

"?" indicates trials on which participants make a response without feedback

strength of the association between the imperfect cue and the common outcome.

Combined (ABC) trials are a combination of the first two transfer trials. These trials also tend to elicit a bias towards the common outcome, but often to a lesser extent than that on imperfect trials (e.g., Johansen, Fouquet, & Shanks, 2010; Kruschke, 1996; Kruschke, 2001a, 2001b; Medin & Edelson, 1988; Shanks, 1992; Winman, Wennerholm, Juslin, & Shanks, 2005; Wood & Blair, 2011), and sometimes show no bias, or even a slight rare bias (e.g., Bohil et al., 2005; Kruschke, 2001a, 2001b; Lamberts & Kent, 2007; Sherman et al., 2009; Wood, 2009). As such, combined test trials show response biases that are less reliable than biases on imperfect and conflicting trials.

Other transfer trials of interest are the perfect predictors, which were only associated with one outcome during training. Comparing accuracy on perfect common (B) and perfect rare (C) predictor trials at test has been highlighted as important for attention-based cue competition accounts of the effect (Le Pelley, Mitchell, Beesley, George, & Wills, 2016; Wills, Lavric, Hemmings, & Surrey, 2014; see *Attention accounts* section below).

Why is the inverse base-rate effect important?

The inverse base-rate effect is generally considered an irrational choice bias and, if it is symptomatic of decisions that we make in everyday life, then the psychological processes responsible for the effect may well have important real-world consequences. Base-rate neglect has been shown to result in an overestimation of disease likelihood in medical professionals (e.g., Casscells, Schoenberger, & Graboys, 1978). Given the potential implications for misuse of base-rate information, it is important to understand why the inverse base-rate effect occurs, and the processes that are responsible for these kinds of biases. As we discuss later, research on the inverse base-rate effect may be valuable not only because of what it can tell us about the information people tend to rely on when faced with conflicting or ambiguous information, but also what it can reveal about fundamental learning processes. Seemingly irrational biases often arise as a product of generally adaptive processes (Tversky & Kahneman, 1974). However, the inverse base-rate effect is also a demonstration of a bias that feedback learning does not seem to correct. Rather, it appears to arise as a result of experience with trial-and-error learning with feedback. There are few category learning effects that involve this kind of deviation from a "normative" standard. Investigating the mechanisms underlying these effects can therefore inform our understanding of the processes that drive human learning and decision-making more generally.

However, despite this potential import and despite occasional robust theoretical debate (Kruschke, 2003; Winman, Wennerholm, & Juslin, 2003), the effect has not had the impact that one might expect, either on informing debates about base-rate neglect as a general property of human cognition or informing relevant theories of learning, including associative accounts of contingency learning. This reflects some uncertainty about the relevance of the phenomenon to general cognitive processes. Indeed, Winman et al. (2005, p.812) argued that the inverse base-rate effect is simply "yet another example of behaviour by puzzled participants trying to figure out what to do in a contrived experimental dilemma". As we discuss in this review, some authors have argued that the inverse base-rate effect is the product of specific inferential reasoning processes that are made in the unique situation posed by the conflicting BC trials rather than being the product of more general learning and memory mechanisms. It is important to note that the critical issue when it comes to the relevance of the inverse base-rate effect is not whether learning and decisions in these tasks are inferential or associative in nature, nor indeed whether they are rational or irrational. Instead, the issue is the assumed specificity of the explanation, and its lack of generalizability to causal and category learning in other contexts. If the decisions that give rise to the inverse base-rate effect are highly idiosyncratic, reliant on a unique configuration of circumstances and thoughts (as it may be argued is the case for some explanations of the effect, such as the eliminative inference discussed below), then it is possible that the phenomenon tells us very little about the rest of our decisions made in other contexts, including the propensity for base-rate use and neglect in learning from experience. In contrast, if the effect is the result of more general learning and attention mechanisms, or the result of inferential reasoning processes that are commonplace in human thoughts and actions, then the effect is relevant and should not be ignored.

It is therefore extremely important to consider the plausibility of the specific explanations that have been offered for the inverse base-rate effect, as well as their implications for what the effect means in the broader context of human cognition. To this end, we offer critical evaluations of some of the more prominent explanations of the effect and highlight where the evidence that is held to be supportive of these explanations is currently relatively weak. We argue that there is still much work to be done to test the generality of the inverse base-rate effect, as well as its underlying causes.

Notwithstanding this uncertainty, the attention that the inverse base-rate effect *has* received reveals it to be of potential import to theory development. One reason the effect was initially considered striking is because it was not predicted by existing exemplar-based models of category learning, such as Medin and Schaffer's (1978) context theory, which anticipates consistent use of base-rates, and other connectionist models using a prediction error or "delta" rule (e.g., Medin & Edelson, 1988; Rescorla & Wagner, 1972; Rosenblatt, 1961; Shanks, 1992; Widrow & Hoff, 1960). The effect has therefore also been important for our understanding of *cue competition*, that is, how predictive signals appear to compete for learning. Cue competition, as a theoretical construct, and its associated learning phenomena have formed the impetus for a whole field of learning algorithms and computational analysis over the past 40 years. The inverse base-rate effect shares important similarities in design and potentially psychological processes with other well-known cue competition effects, such as blocking. It is also highly related to the learned predictiveness effect (or learned predictiveness principle; Le Pelley & McLaren, 2003; Lochmann & Wills, 2003; Mackintosh, 1975), in which predictive cues capture greater attention than less predictive cues, and are consequently learned about more readily. Kruschke (2001a, 2003) has argued the effect is functionally similar to *highlighting*, in which AB-O1 compounds are first learned in an initial training phase, prior to the introduction of AC-O2 trials. In the highlighting effect, BC trials elicit a preference for the late outcome, O2. While research on the inverse base-rate effect was most prominent between the late-1980s and early-2000s, it has recently received renewed interest (e.g., Don & Livesey, 2017; Don, Beesley, & Livesey, 2019a; Inkster, Milton, Edmunds, Benattayallah, & Wills, 2019a; Inkster, Mitchell, Schlegelmilch, & Wills, 2019b; Le Pelley et al., 2016; O'Bryan, Worthy, Livesey, & Davis, 2018; Wills et al., 2014). Specifically, for reasons that are discussed below, the inverse base-rate effect has been highlighted as an important phenomenon because it may discriminate between different attention-based models of learning. Further, the effect provides a useful opportunity to examine how higher-order reasoning processes may interact with lower-level processes in learning and decision-making. In the following sections, we examine the strengths and weaknesses of common theoretical explanations for the inverse base-rate effect, and highlight important remaining questions that should be addressed in future research. As a resource for researchers interested in studying the inverse base-rate effect, we have also included a summary of the methods typically used to measure the effect.

Theoretical accounts of the inverse base-rate effect

A novelty effect

Perhaps the simplest and most intuitive explanation for rare outcome biases is in terms of a relative novelty effect (Binder & Estes, 1966), which combines the idea that novel or striking events are more memorable (*Rhetorica ad Herennium*, c.85BC) with the *availability heuristic*, which states that events more easily remembered are judged to be more probable (Tversky & Kahneman, 1973). The availability heuristic is, in essence, an associative principle. That is, the stronger the association between a cue and an outcome, the greater the ease with which a cue will "bring to mind" an outcome (Hamilton, 1981). However, a novelty explanation in its simplest form is disconfirmed by the observation that imperfect (A) and combined (ABC) transfer trials tend to elicit responses consistent with the underlying base-rates (Medin & Edelson, 1988). A relative novelty explanation instead predicts that these trials would also bring to mind the rare outcome. Further, compounds trained in the same base-rates but without a shared cue elicit base-rate normative responding for conflicting cues, which also suggests the effect is not a bias based on the novelty of the cues or outcomes (Kruschke, 2001a; Medin & Edelson, 1988; Wills et al., 2014).

Associative learning and cue competition

According to associative learning theories, contingency learning results from the formation and strengthening of associations between cues and outcomes. The relationship between a cue and an outcome depends not only on their co-occurrence, but also the predictive qualities of other cues presented at the same time. According to associative accounts, selective learning effects arise due to simultaneously presented cues competing for a limited amount of associative strength with the outcome (Dickinson, Shanks, & Evenden, 1984).

Several prominent cue competition effects are well explained by prediction error models of learning. In these models, learning only occurs to the extent to which an outcome is surprising, or unexpected (Kamin, 1969; Rescorla & Wagner, 1972). Thus, learning about a cue can be restricted by the presence of another that already predicts the outcome. Perhaps the most widely cited formalisation of this idea is the Rescorla-Wagner model (Rescorla & Wagner, 1972), though the same concept has been used in many other models of predictive learning.

In inverse base-rate tasks, predictive cues are trained in compound with an imperfect predictor. It is therefore possible that differential competition amongst cues in common and rare compounds may result in differences in learning about predictive cues. Because cue A is a better predictor of O1 than O2, it should compete more effectively with cue B than with cue C for associative strength with their respective outcomes. This would then result in a weaker association between B and O1 than between C and O2, such that C controls responding on BC trials. Indeed, some authors have noted negative correlations between responding for A and responding for BC trials at test (Medin & Edelson, 1988; Shanks, 1992). That is, the more strongly A is associated with O1, the more C appears to dominate responding on BC trials. However, several authors have demonstrated that the Rescorla-Wagner model is unable to account for the inverse base-rate effect (Gluck & Bower, 1988; Markman, 1989). As cue A is an imperfect predictor, it eventually loses associative strength,[3] while predictive cues B and C gain all associative strength with their respective outcomes. Learning about AB also reaches asymptote more quickly than learning about AC, due to the difference in presentation frequency, and therefore prior to asymptote, B will always be more strongly associated with O1 than C is with O2. Consequently, if BC is tested at asymptote, Rescorla-Wagner predicts no bias in choice, whereas if BC is tested early in training, the model predicts a common outcome bias. Thus, there is no point at which the model predicts a rare bias for BC trials.

Several authors have proposed additional model assumptions that may allow Rescorla-Wagner to account for the effect. For example, Markman (1989) suggested that if the activation of absent cues is coded as -1, rather than zero, the model could predict an inverse base-rate effect. However, as Shanks (1992) pointed out, it is difficult to determine which of all possible encountered cues should be considered absent, and so this assumption may prove problematic in practice (but see Dickinson & Burke, 1996; Larkin, Aitken, & Dickinson, 1998; Van Hamme & Wasserman, 1994, for potential solutions). Gluck (1992) proposed that incorporating distributed cue representations could account for the effect, but this only predicted a small preference for the rare outcome on BC trials, and a small preference for the common outcome on A and ABC trials. In practice, the biases on BC and A trials are much more substantial. In addition, Gluck's model does not predict greater accuracy for B than C at test that is found in some studies (e.g., Wills et al., 2014 – described in greater detail in the following section).

[3] Shanks (1992) notes that A will be a better predictor of the outcomes than a neutral cue, and therefore should not lose *all* associative strength.

Attention accounts

Although several explanations of the inverse base-rate effect have made use of connectionist models (Gluck & Bower, 1988; Medin & Edelson, 1988; Shanks, 1992), the most widely accepted account is the attention-based approach proposed by Kruschke (1996, 2001b). *Attention* refers to the mechanisms responsible for prioritising certain stimuli or events for further processing. Attention-based theories of associative learning (e.g. Kruschke, 2001a, 2001b; Mackintosh, 1975; Pearce & Hall, 1980) posit that attention is flexible and is influenced not only by cue salience but also by previous experience of the relationship between cues and outcomes. According to these theories, processes of learning and attention interact. That is, learning about the relationship between a cue and an outcome determines the amount of attention allocated to a cue, and the amount of attention allocated to a cue influences the rate of learning about that cue in future associations. It is widely established that cues that are reliable predictors of outcomes attract preferential attention, with extensive evidence in animal learning (Mackintosh, 1975; Sutherland & Mackintosh, 1971) and human learning (see Le Pelley et al., 2016, for a review).

According to the attention-based account, choice of the rare outcome on conflicting BC trials is explained as the result of increased attention to cue C during learning (Kruschke, 1996, 2001b). AB-O1 trials are learned well early in training, because they occur more frequently than AC-O2 trials. As a result, both A and B form associations with the common outcome, to some extent. As this learning has primacy, the presence of A on AC trials quickly comes to produce an incorrect prediction of O1. Kruschke's model assumes a shift in attention occurs to reduce this error. At the point where prediction error is experienced, attention rapidly shifts away from the ambiguous cue A towards the more predictive cue, C. This kind of attention towards more predictive cues forms the basis of the learned predictiveness principle (Le Pelley & McLaren, 2003; Lochmann & Wills, 2003). The attention shift serves the dual purpose of correcting error on subsequent AC trials more effectively and also preserving what has been learned about AB trials. The resulting attention bias to C supports a stronger association between C and O2 than the association between B and O1, and may also produce continued attention to cue C on BC trials, such that C tends to control responding on BC trials at test. Either of these two biases – one based in learning, the other in test performance – could independently produce a tendency to choose the rare outcome when presented with BC for the first time on test.

Evidence for attention accounts

Several studies have provided evidence in support of the role of attention in producing the rare outcome bias on conflicting trials.

The importance of the imperfect predictor The necessity of a shared cue (e.g., cue A is present in both the common and rare compounds) provides particularly compelling evidence for accounts of the effect that are mediated by prediction error, such as the attentional account described above. Several studies have shown that the inverse base-rate effect does not occur for conflicting DE trials if D and E are trained in non-overlapping compounds, where FD-O1 is trained more frequently than GE-O2 (Kruschke, 2001a; Medin & Edelson, 1988; Wills et al., 2014). In the absence of a shared cue, there would be no prediction error on GE trials to drive attention toward the predictive cue, E. Wills et al. (2014) measured event-related potentials (ERPs) associated with visual attention in response to predictive cues after training with a shared cue (AB vs. AC) and after training in the absence of a shared cue (FD vs. GE). This included differential ERP effects, Selection Negativity and Selection Positivity, which indicate the difference in ERPs for the target and unattended stimuli, and are elicited by attention to features. When cues were presented individually at test, there were significantly greater posterior Selection Negativity and concurrent anterior Selection Positivity for C compared to B. However, there was no significant difference in these ERPs for E compared to D, which had not been trained with a shared cue, and had not elicited a rare bias when presented in compound at test. This suggests greater attention to the rare predictor, but only when it had been trained in compound with an imperfect predictor, which suggests that error-driven shifts of attention may be critical in driving choice biases.

Eye gaze and overt attention Eye gaze is often used as a measure of overt attention. While it is possible to make covert shifts of attention without accompanying eye movements, spatial allocation of attention and gaze direction are generally closely related (Posner, 1980). Kruschke, Kappenman and Hetrick (2005) trained participants in a highlighting design (where AB-O1 is trained in a first phase before the introduction of AC-O2 in a second phase), while measuring fixation time (the length of time spent fixating on each cue). They found greater fixation time to C on AC trials than to B on AB trials, and greater fixation time to C than to B on BC trials. In a standard inverse base-rate design, Don et al. (2019a) found greater fixation time to C than A on AC trials during training, and no bias on AB trials, both prior to making a choice prediction, and during feedback. Prior to making a decision, the attention bias towards C increased across the course of training, likely reflecting a learned attention bias towards cues that best predict the outcome. During feedback, the bias towards C emerged quickly and reduced over the course of training. This likely reflects an attention shift in response to error, with attention directed towards cues that will reduce future error, and subsequently decreases as prediction accuracy increases.

Attention transfer Most models based on predictiveness principles (e.g. Kruschke, 2001b; Le Pelley, 2004; Mackintosh, 1975; Pearce & Mackintosh, 2010) assume that participants attend to cues according to their predictive history, and that learning about cues is proportional to the attention allocated to them. Attention to cues influences their associability, or the rate at which they are learned about. That is, the more attention paid to a cue, the faster that cue will be learned about in future novel associations. In a recent study, we show greater associability for previously rare predictors (C) than previously common predictors (B; Don & Livesey, 2021). After base-rate training, in a new learning phase, previously rare predictors and previously common predictors were presented in compound, and paired with a novel outcome. Participants learned the association between previously rare predictors and the novel outcome better than the association between previously common predictors and the novel outcome. This transfer bias was evident after a short amount of base-rate training but was diminished after longer training. A similar result has been reported for highlighting. Kruschke et al. (2005) gave participants a secondary training phase with several new combinations of either AB or AC trials, paired with novel outcomes. In this phase, A was now predictive of the novel outcome, while B and C were non-predictive. If C captures more attention than B during training, and this continues into future learning, it should be more difficult to learn about the new predictive cue A in the presence of C than in the presence of B. Indeed, Kruschke found that AC trials were learned more slowly than AB trials, between subjects, thereby providing an indirect test of the difference in attention for C versus B.

Salience effects An assumption of the learned predictiveness principle, and attentional shifts that may be responsible for the inverse base-rate effect, is that predictiveness operates a little like salience, making cues that have been predictive stand out in the same way as physically salient stimuli. More salient stimuli also attract greater attention (Denton & Kruschke, 2006). Cue salience can be determined not only by the physical properties of the cue, but also training history (Mitchell & Le Pelley, 2010). In a medical diagnosis task, Bohil, Markman and Maddox (Bohil et al., 2005) tested the hypothesis that cue C gains greater salience and attention throughout training by manipulating the perceived salience of cues, while training AB-O1 and AC-O2 trials in equal base-rates. They found that if cue C was presented as a serious symptom (e.g., paralysis), and cue B was presented as a mild symptom (e.g., stuffy nose), participants showed a preference for O2 on BC trials, creating an analogue of the inverse base-rate effect. They also found that when AB-O1 and AC-O2 were trained in a 3:1 base-rate, the inverse base-rate effect could be removed if cue B was presented as a more serious symptom than cue C. If similar processes also govern the inverse base-rate effect,

these results imply the rare predictor develops greater salience as a result of training, and therefore receives greater attention.

Attention models and the inverse base-rate effect

The inverse base-rate effect provides a good test-bed for models of attention in learning. The attention account of the inverse base-rate effect was initially formalised in the ADIT model, which was later extended to allow attention distributions to be learned in the EXIT model (Kruschke, 1996, 2001b). These models have been successful in predicting the inverse base-rate effect and related highlighting effects (Kruschke, 2001a, 2003). The EXIT model is a connectionist model of associative learning that incorporates rapid attention shifting in response to prediction error. That is, after making an outcome prediction, and subsequently observing the actual outcome, attention shifts to the cue most likely to reduce subsequent prediction error. After attention shifts within a trial, the distribution of attention to each cue is incrementally learned, such that a proportion of that attention distribution can be applied on similar trials in the future. In EXIT, attention moderates the influence of cues on both responding and learning, such that cues that are attended will have greater control over responding within a trial, and will be learned about more readily. Thus, according to EXIT, attention is shifted towards C on AC trials, due to the prediction error driven by the ambiguous cue A. On BC trials, C has a stronger association with O2, and/or greater attention is paid to cue C, resulting in greater O2 responses. EXIT is a relatively complex model containing several mechanisms. Recently, 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, including the inverse base-rate effect.

While the EXIT model is most commonly applied to the inverse base-rate effect, it is not the only attention-based model of learning that is relevant to the effect. EXIT is based on the theoretical principle that greater attention is allocated to cues that reduce subsequent prediction error, and is therefore conceptually very similar to the Mackintosh (1975) model. The Mackintosh model has been applied extensively in animal and human learning literature, and has been critical in explaining related biases in learning, such as the learned predictiveness effect (e.g., Le Pelley et al., 2016; Le Pelley & McLaren, 2003; Lochmann & Wills, 2003). While the Mackintosh model can account for choice biases in the inverse base-rate effect, it fails to predict stronger attention to C on AC trials than to B on AB trials (Don et al., 2019a), as well as associability benefits for previously rare predictors (Don & Livesey, 2021).

More recently, Le Pelley et al. (2016) proposed a far simpler model in which attention is assumed to be proportional to associative strength. Here, attention would be directed towards cues that are good predictors of an outcome, and away from cues that may weakly predict multiple outcomes. The authors state that this model can account for most effects of attention in human learning, with the exception of the inverse base-rate effect. The primary issue for this model is the finding that often, the inverse base-rate effect occurs when common responses for B alone are greater than rare responses for C alone at test (e.g., Inkster et al., 2019b; Wills et al., 2014). We refer to this difference as the B>C effect. This suggests that the inverse base-rate effect occurred even though there was a greater association between B and O1 than between C and O2. While the co-occurrence of the inverse base-rate effect and the B>C effect is not always present, this trend is evident in several other cases (e.g., Bohil et al., 2005; Kruschke, 1996; Medin & Edelson, 1988; Medin & Bettger, 1991; Shanks, 1992; Winman et al., 2005; see Winman et al., 2003, for further discussion of this issue). The result is surprising, as rare responding to BC trials cannot be predicted by the summation of associative strengths for B and C. The finding may be problematic for attention-based theories of the inverse base-rate effect, particularly for the simple attention account that assumes a close relationship between learning strength and attention biases (Le Pelley et al., 2016). At the very least, it suggests that there may be dissociations between attention and response accuracy (Wills et al., 2014; Winman et al., 2003).

Context associations

One potential way to reconcile the inverse base-rate effect and the B>C effect with Le Pelley et al.'s attention model is to assume a role of context learning (Don & Livesey, 2017; Don et al., 2019a; Le Pelley et al., 2016). Context associations are an important component of associative learning models for a range of reasons, but, in this case, particularly because they are the primary way in which such models can explain learning of base rates that is independent of predictive cues. The assumption is that the experimental context acts as an additional cue that becomes associated with the outcomes, and influences predictions and judgments in much the same way as other cues. Because of the base-rates, the context will be more strongly associated with the common outcome. Therefore, B alone trials can be considered B+context trials, where both the cue and context associations would facilitate prediction of the common outcome. On C+context trials, the cue and context predict different outcomes, such that context associations might weaken rare outcome predictions, even if the C-O2 association is stronger than the B-O1 association.

The effect of context associations has been studied by equating global outcome frequency (Don & Livesey, 2017; Don et al., 2019a, Don & Livesey, 2021). In typical inverse base-rate designs with multiple instantiations and fewer

outcomes than cue compounds, each outcome is always paired with either common compounds or rare compounds. For example, O1 is always paired with common compounds AB and DE, and O2 is always paired with rare compounds AC and DF, such that the context will be more strongly associated with O1. Don and colleagues used this as the "standard" training condition. In contrast, the experimental condition in these studies was a "balanced" training condition in which each outcome was paired with both a common compound and a rare compound. In this way, all outcomes were experienced with equal frequency during training. For example, O1 was paired with the common compound AB and the rare compound DF, while O2 was paired with the rare AC and the common DE. For any given set of overlapping cues, the base-rate difference is preserved – O1 was the common outcome within overlapping AB and AC trials, and O2 was the common outcome within DE and DF trials – and these *local* base-rate differences ensure that there is a rare predictor and common predictor in each set. At the same time, there is no *global* base-rate difference in the frequency of the outcomes, and thus there is no basis for the context to be more strongly associated with any particular outcome.

The balanced outcome design typically reduces the preference for the rare outcome on conflicting trials (Don & Livesey, 2017; Don et al., 2019a). Differences in the B>C effect as a result of context associations were difficult to assess in these studies, as responding for both B and C test trials were close to ceiling in both groups, such that no difference was observed. Don et al. (2019a) also found that global outcome frequency produces differences in gaze preferences prior to making an outcome prediction during training. While the standard group in their study showed preferential attention to the predictive cue on AC trials, and no bias on AB trials, the balanced condition produced gaze biases towards the more predictive cue on both AB and AC trials. As the context does not provide a good prediction of the outcome, it may be more necessary to attend to the predictive cue on every trial in order to make an accurate prediction. In contrast, the standard group can rely to some degree on the global base-rates to make correct predictions, with the exception of AC trials. During feedback, where, according to EXIT, attention is adjusted to minimise prediction error by attending more to the cues that produce the least error, there was no difference in attention between standard and balanced groups, perhaps suggesting that regardless of the role played by context learning, attention tends to shift to discrete cues during this phase. Assessing attention during feedback is not common practice in these types of tasks, and therefore warrants further research.

In the EXIT model, context learning is captured by associations between the outcome and a "bias node", which is allowed to vary in salience. The EXIT model can predict the co-occurrence of the inverse base-rate effect and B>C effect, but only when the B versus C difference is heavily weighted in model fits (Kruschke, 2001a, 2003). Notably, EXIT can account for reductions in the inverse base-rate effect and attention biases when outcome frequency is matched (Don et al., 2019a).

While the inverse base-rate effect is reduced when the context is not strongly associated with the common outcome, it is unlikely to be the result of context conditioning alone. In one study (Don & Livesey, 2017, Experiment 3), AB-O1 and AC-O2 were trained in equal frequency, and outcome base-rate differences were produced by including high-frequency filler trials paired with O1, such that there should be a strong context-O1 association. These conditions were not sufficient to produce an inverse base-rate effect on BC trials. In addition, Inkster et al. (2019b) found that changing the context in the test phase had no effect on the strength of the inverse base-rate effect or B>C effect. The failure to find any effect of a context shift on test performance is not consistent with the notion that context associations are heavily involved in the inverse base-rate effect. It is clear there is a need for further research to identify the role of context in the inverse base-rate effect, and its implications for models of attention.

Inferential accounts

Some researchers have proposed that, rather than a learning effect, the inverse base-rate effect is a rational decision based on inferential processes at test (Juslin, Wennerholm, & Winman, Juslin, Wennerholm, & Winman, 2001; Winman et al., 2005). According to this account, selective attention and selective learning are not necessary to explain rare outcome choice. The predominant inferential explanation is the *eliminative inference* account, proposed by Juslin et al. (2001). The general principle of this account is that people eliminate the well-known, common category when they are faced with ambiguous or dissimilar features that do not fit that category. For instance, BC trials do not match the known AB features of the common outcome, and therefore that option is eliminated, and the rare outcome is instead chosen.

More formally, the eliminative inference model (ELMO) assumes that participants form a number of inference rules about the relationship between cues and outcomes (e.g. AB → O1) during training. On any given trial during the test phase, some of these learned rules will form part of an *active set* in working memory. The remaining outcomes, which pertain to inference rules that are not currently active in memory, form part of a *guessing set*. The probability of an inference rule being part of the active set is proportional to its base-rate, so that frequent inference rules (e.g. AB → O1) are more likely to be active in memory than infrequent inference rules (e.g., AC → O2). At test, ELMO determines whether a particular transfer item will elicit a process of *induction* or *elimination* according to the similarity between each of the

active inference rules and the transfer item, where similarity is determined by the cues (or features) comprising the inference rule and the transfer item. Transfer items with at most one deviating feature to the active rules will elicit induction. In this case, inference rules that share similarity with the transfer item are activated, and an active outcome is then chosen at a probability proportional to its similarity to the transfer item. For example, A and ABC transfer trials have one dissimilar feature to the active rules (missing one perfect predictor, or one too many perfect predictors, respectively), and therefore lead to induction. Because the common inference rule is more likely to be part of the active set, O1 is typically chosen for these trials. Transfer items with two or more deviating features will instead elicit elimination. In elimination, the active inference rules (and thus the choice of the relevant outcome) are removed, and participants guess amongst the remaining outcomes (i.e. those whose inference rules are inactive) in the guessing set, which is more likely to contain rare outcome rules. The conflicting BC trial deviates from AB → O1 with two features (one missing imperfect predictor, and one extra perfect predictor), and so the active O1 outcome is eliminated, and participants guess the rare O2 outcome. As is evident from this model, an eliminative inference is highly specific in terms of the rules surrounding induction and elimination, which may only apply in an artificial learning environment when making discrete choices. It is difficult to see how these specific rules may generalise readily to other situations in everyday life or even other similar phenomena in the lab. However, elimination in general may be a cognitive mechanism that is more widely used. We discuss this further later in the paper.

Although ELMO accounts for the inverse base-rate effect in its original form, the model fails to account for several important characteristics of the effect. These include the necessity of a shared cue during training, as ELMO incorrectly predicts elimination even in cases where there is no imperfect predictor (Wills et al., 2014). The model also struggles to account for differences in the strength of responding to imperfect and combined transfer items; typically, the tendency to choose the common outcome is stronger for A than ABC trials, even though ELMO predicts induction should occur for both (Kruschke, 2001a). The model also incorrectly predicts elimination on transfer trials that combine an imperfect predictor with a novel cue. For instance, Don and Livesey (2017) found that AX trials elicited a tendency to choose the common outcome, even though these trials differ from the AB → O1 rule by the same number of features as the BC transfer trials. As such, many researchers no longer consider it a plausible model of the effect.

The way in which ELMO is defined, operating on a set of discrete rules (one for each outcome category or training trial) does not lend itself easily to some of the conditions under which the inverse base-rate effect is readily observed, for instance when multiple cue combinations lead to the same outcome (and indeed the fact that this has no effect on rare bias, see Don & Livesey, 2017, Experiment 1) or when the cues are not discrete but presented as variable quantities (Kalish, 2001). These properties could at least, in principle, be accommodated by a generalised version of the model. Potentially more consistent with a decision process based on discrete rules, Kalish (2001) did not observe the effect when using overlapping distributions of cue values, which in effect produced probabilistic rather than deterministic cue-outcome relationships. There are several unusual features of this experiment that may have led to this failure but if it were the case that probabilistic relationships did not support the inverse base-rate effect then that would be rather inconsistent with a connectionist account. Again, it would also cast serious doubt on the generality of the phenomenon in question. However, more research is needed to answer this question one way or another.

Although there are serious doubts about the ability of eliminative inference to adequately account for the inverse base-rate effect, this does not mean some form of reasoning process at test is not involved in producing it. Some researchers have suggested that elimination is synonymous with a general novelty-matching strategy, in which novel cues are paired with novel outcomes. While we suggest this kind of matching strategy differs from a dissimilarity-based eliminative inference (the matching strategy involves selecting the rare outcome specifically because of its novelty rather than eliminating the common outcome and choosing among the rest), this is at least another reasoning process that is not captured by connectionist networks like the EXIT model (Kruschke, 2001a). The following section outlines evidence related to higher-order reasoning processes in the effect.

Evidence for inferential accounts

Novel cue effects

Responding to novel cues has been an important test for elimination accounts. That is, a novel cue at test should elicit elimination of the common outcome, due to its dissimilarity to learned rules. Juslin et al. (2001) found a rare outcome bias for novel cues, and Johansen, Fouquet, and Shanks (2007) found a similar result in a text-based version of the task. However, Inkster et al. (2019b) found a common bias on novel cue trials, and, as noted, Don and Livesey (2017) found a bias for the common outcomes on AX trials that contained the imperfect predictor plus a novel cue, which should also elicit a process of elimination. In this case, responding to novel cues may be based on the associations of other cues present (e.g., context, imperfect predictors).

Processing of the common predictor

In an fMRI study, O'Bryan et al. (2018) used multivoxel pattern analysis to examine activation patterns associated with the cues presented on ambiguous test trials. Faces, objects and scenes were used as cues, as these visual categories have well-defined regions of representation in ventral temporal cortex. Prior to training, regions of sensitivity to these categories were determined for each participant. The authors used representational similarity analysis to compare patterns of activation from this initial task with patterns of activation on conflicting test trials, giving an indication of whether participants are more strongly activating information from the common or the rare category. Although there was no significant overall preference for the rare outcome, on conflicting trials where the rare outcome was chosen, there was neural activity indicative of stronger activation of cue B than of cue C. No such difference emerged on trials where participants chose the common outcome. This finding suggests that participants were choosing the rare outcome by making a decision about the *dissimilarity* between the test trial and the common category trials (AB – O1) experienced during training, therefore avoiding the common outcome as a choice. O'Bryan et al. interpreted this as evidence of a deliberative form of reasoning, consistent with the eliminative inference, and distinct from similarity-based choices. Consistent with this interpretation, response times were also slower on conflicting trials when rare choices were made, but were faster when the neural activity indicated the common outcome was processed to a greater extent. In other words, participants were faster to respond with the rare outcome if they attended more to the common cue B. However, another study has shown greater activation for cue C than cue B in brain regions linked to prediction error (Inkster et al., 2019a).

Developmental and comparative differences in base-rate adherence

Researchers often turn to developmental changes and cross-species differences in behavioural phenomena in an attempt to understand their psychological origins, and to some extent this is true for the inverse base-rate effect. Winman, Wennerholm, Juslin and Shanks (Winman et al., 2005) compared the effect between adults and children, assuming that children would be less likely to use high-level reasoning processes, and instead rely on more simple associative processes. Adults demonstrated an inverse base-rate effect, while children showed responding at chance levels on conflicting trials. The authors argue that the difference between adults and children was not attributable to differences in learning efficiency, suggesting that the result was not simply due to differences in the strength of learned associations or selective attention (since these should manifest in differential performance on the task). The authors thus argued that the effect observed in adults was likely to be a result of their capacity to use higher-order processes. However, in contrast to these results, Burling and Yoshida (2016) found highlighting effects in young children.

To date, the inverse base-rate effect and highlighting effect have only been demonstrated in humans. In comparison, some cue competition effects such as blocking and overshadowing have been found in many other species, using a wide variety of conditioning tasks. In the only comparative study of the highlighting effect that we know of, Fagot, Kruschke, Depy and Vauclair (1998) compared the effect in humans and baboons. While humans showed a rare bias on conflicting trials, baboons showed ambivalent responding on both conflicting and imperfect test trials. Fagot et al. argued that this was due to a difference in rapid attention shifting, rather than reasoning capacity. However, it is clear that if the effect in humans was based on higher-order reasoning then we might expect to see a weaker or even absent effect in other species. In any case, since the only data to date that bear on this question come from just two baboons, comparative evidence for or against the inverse base-rate effect is sorely lacking. Future studies could test the effect in animals using simple conditioning paradigms.

Individual differences in rule use

Recent studies in category learning have found that individual differences in the tendency to rely on feature- versus rule-based processes are relatively stable across tasks, where "feature-based" refers to generalisation on the basis of surface similarity of features, and "rule-based" refers to generalisation on the basis of abstract relations (Little & McDaniel, 2015; McDaniel, Cahill, Robbins, & Wiener, 2014). This has given rise to the suggestion that certain types of generalisation are often carried by a subset of participants who have a tendency to search for abstract relations among items (Goldwater, Don, Krusche, & Livesey, 2018). If only a subset of participants (e.g., those disposed to rule-learning) exhibited the inverse base-rate effect, then it might suggest the effect is grounded in selective cognition. To test this hypothesis, Winman et al. (2005) classified participants as rule-based or feature-based, according to generalisation performance in a *patterning* task (Shanks & Darby, 1998), in which rule-transfer and feature-based transfer produce very distinct patterns of generalization on test. Other studies have found that rule-based generalisation in this task is associated with higher working-memory capacity (Wills, Barrasin, & McLaren, 2011a), greater cognitive reflection, and more strategic "model-based" choice in reinforcement learning (Don, Goldwater, Otto, & Livesey, 2016). In this case, Winman et al. (2005) compared the strength of the inverse base-rate effect produced by participants classified as rule- or feature-learners, based on their performance in a separate patterning task. Only those

participants who were able to extract and apply an abstract rule in the patterning task exhibited an inverse base-rate effect, suggesting that the effect may rely on the use of higher-order processing. Nevertheless, this observation is open to multiple interpretations, particularly because it is correlational in nature. For instance, there is evidence of individual differences in selective attention in tasks widely assumed to have their basis in broad learning mechanisms (e.g., Granger, Moran, Buckley, & Haselgrove, 2016; Haselgrove et al., 2016; Le Pelley, Schmidt-Hansen, Harris, Lunter, & Morris, 2010), and this may form the basis of individual differences in the inverse base-rate effect. People who display rule-based learning in the patterning task typically also display greater learning efficiency (i.e., they improve faster) and faster learning may provide an advantage for learning rare predictors in particular since they are less frequent and typically the slowest to attain high accuracy. While Winman et al. (2005) attempted to control for overall training accuracy in their statistical analysis, there is still a possibility that rule-learning in the patterning task and the strength of the inverse base-rate effect are indirectly linked by a common but very general factor such as participants' attentiveness or motivation to engage on the task, or that each is the result of different cognitive processes that rely on a common cognitive capacity, as would be the case for instance if some component of attention shifting were working memory dependent. Without knowing the basis of this link, we can only conclude that people who are likely to pick up and use the patterning rule are also likely to show the inverse base-rate effect, but the reason for this association is far from clear.

Is learning through experience necessary?

Johansen et al. (2007) demonstrated an inverse base-rate effect when cue-outcome contingencies were presented simultaneously in written listed format, suggesting that trial-by-trial learning is not necessary for the effect. Although the rare choice bias was present when each trial was listed in text form, the effect was negated when an explicit summary of the base-rates was provided, suggesting the choice bias is not simply based on higher-order processes applied at test. Curiously, in the inverse base-rate effect, summarising base-rates appears to encourage their use (e.g., Johansen et al., 2007), yet the opposite appears to be true for base-rate neglect (e.g., Butt, 1988; Christensen-Szalanski & Beach, 1982; Christensen-Szalanski & Bushyhead, 1981; Manis et al., 1980). However, the task requirements involved in producing these phenomena are very different.

Cognitive load

Lamberts and Kent (2007) argued that there is no evidence for rule-based processes in the inverse base-rate effect. In their study, participants were tested in four different within-subjects conditions, three of which were designed to tax working memory capacity, and therefore interfere with effortful, cognitively demanding reasoning processes. In a control condition, participants responded to test trials under standard, unspeeded conditions. In a dual-task condition, participants were required to simultaneously count backwards in multiples of three, and in two speeded conditions, participants were required to respond to test trials within either 500 or 300 ms. An inverse base-rate effect was obtained in all four conditions. The authors therefore argued that the inverse base-rate effect cannot be a result of inferential processes at the time of test.

Some aspects of Lamberts and Kent's (2007) analysis make it difficult to draw a firm conclusion that these conditions have no influence on the strength of the effect. For instance, differences in the strength of the effect under dual-task or speeded conditions, compared to no-load conditions, were not reported. Indeed, the effect appears numerically weaker under speeded conditions. Further, the cognitive load manipulations were applied during the test phase only. The assumption is that inferential rules in the inverse base-rate effect are formed or used effortfully during the test phase, because this is where the relevant trials are presented. Nevertheless, it is possible that rules or inferences are formed throughout training, and these inferences may then be applied during the test phase with little cognitive demand. Indeed, Wills, Graham, Koh, McLaren and Rolland (2011b) found that the use of rule-based processes in a patterning task was affected by cognitive load during training, but not during test. Rule-use in the patterning task is also based on responses to new combinations of trained cues, in a similar way as the inverse base-rate effect. Thus, Lamberts and Kent's (2007) results do not rule out the possibility that some form of inferential reasoning could contribute to the effect.

Related to this, the inverse base-rate effect is affected by conditions that may influence the way in which inferences are formed during training. That is, the effect is weakened when outcome novelty or trial novelty is removed (Don & Livesey, 2017). However, whether the effect of novelty influences conscious reasoning processes or lower-level learning processes (or both) is unclear.

Self-report measures

The ability to articulate a rule is often considered an important part of inferential rule use and, indeed, in tasks where a dominant and clearly articulable rule is in place, articulation of the rule is often found to be highly correlated with patterns of generalization (e.g., Lee, Hayes, & Lovibond, 2018). To the best of our knowledge, no study has attempted to systematically document participants' articulated response rules in relation to the inverse base-rate effect.

Table 2 Outcome choice in the questionnaire by bias shown at test

Bias at test	Questionnaire response		
	Common	Rare	Total
Common	3	0	3
Rare	15	4	19
Unbiased	5	3	8
Total	23	7	30

By way of a preliminary investigation, we assessed participants' reasoning processes on conflicting trials by administering a post-experimental questionnaire to 30 participants, after they completed an inverse base-rate effect task using identical methods to the standard condition described in Don and Livesey (2017), which is known to produce a strong inverse base-rate effect, and indeed did so again in this study (mean rare outcome choice on BC trials accounted for 72% of all relevant outcome choices, see Online Supplementary Materials for a full description of methods and learning task results). The aim was to determine whether participants articulate a clear inferential reasoning strategy for rare choice at test, and if so, whether this strategy is consistent across participants. The questionnaire provided a written summary of the abstract structure of the task design, the outcome base-rates, and the conflicting trials. They were asked to indicate which outcome they tended to choose on conflicting trials during the test phase of the computer task, and were then asked to describe their reason for choosing that outcome. Participants were classified as rare-biased, common-biased or unbiased, according to their responses to conflicting trials in the test phase (see Table 2). Table 3 provides category definitions into which outcome choice responses were

Table 3 Explanations for outcome choice by category

Outcome	Category	Classification	N	Mean rare bias	Example response
Common	Base-rate normative	The more frequent outcome has a higher probability of occurring	9	.67 (±.11)	"Since disease 1 occurred more often it makes sense that there is a bigger probability that the patient suffers from disease 1 rather than 2."
	Associative memory	Remembered the relationship between B and O1 better	6	.71 (±.08)	"I remembered the symptoms of this disease better than the other diseases which weren't seen as often."
	Intuitive	Other reason for choosing O1, e.g., greater confidence, or salience of the cue or outcome etc.	4	.75 (±.10)	"I was more confident on the symptoms displayed for disease 1 than 2"; "I would often give symptom C less weighting than B because it would feel somehow more natural for me to do so."
	Unclassified		4	.71 (±.17)	
Rare	Elimination	If the cues do not match those of the common outcome, the rare outcome is chosen.	1	.50	"Disease 1 occurred so often that it was really easy to identify whether or not it was disease 1, and if the symptoms did not match or seem familiar to that of disease 1, I would choose disease 2.
	Novelty Selection				
	Outcome	O2 was chosen because it was a novel outcome	0	-	
	Trial	O2 was chosen because C was novel	2	.50 (±.0)	"…Since symptom C is unusual, when it appears it probably means the patient is more likely to get the rare disease associated with symptom C"
	Novelty-matching	BC was novel, so the more novel outcome was chosen	0		
	Asymmetric representation	C indicates the rare outcome more than B indicates the common outcome.	3	.92 (±.08)	"Symptom C could be the main signal of a rare kind of disease" "Because that symptom seemed to be exclusive to only that disease."
	Unclassified		1	1.0	

Note: Mean rare bias refers to the proportion of relevant rare outcome choices on conflicting BC trials for participants in each classification

classified. As this questionnaire was primarily an exploratory exercise, these categories were determined post hoc, based primarily on proposed explanations for the effect. The number of responses falling into each of these categories, mean rare biases at test for each category, and example responses, are also summarised.

The questionnaire revealed several notable results. Responses in the questionnaire did not clearly correspond with choice at test. Of the 19 participants who were clearly rare-biased at test, 15 indicated they had chosen the common outcome. Although Johansen et al. (2007) also found greater common responses in a similar questionnaire, it is interesting that providing this kind of base-rate summary was also able to override choices made during the test phase. It could be the case that participants simply could not recall which outcome they had chosen on these trials, although the questionnaire was administered directly following the test phase. Alternatively, it may be difficult for participants to map the abstract structure described in the questionnaire to the appropriate trials during the test phase, especially considering the large number of test trials. However, this mismatch in choice did not occur in the opposite direction. That is, no participants who indicated they had chosen the rare outcome in the questionnaire had shown a common bias at test. Thus, it seems to be the case that providing a summary of the base-rates leads to more rational responding, even in participants prone to making rare responses when contingencies are learned on a trial-by-trial basis.

Most participants who chose the common outcome in the questionnaire gave a rational, base-rate normative explanation. That is, the more common outcome is more likely to occur. Several other participants gave an explanation based on associative memory, where they indicated that they remembered the outcome that went with cue B better, because it was seen more frequently. Others relied on more intuitive explanations, such as greater confidence in that response.

The explanations for choosing the rare outcome were of greater interest. One participant clearly articulated an eliminative inference, stating that if the symptoms did not match those of the common disease, they chose the rare disease. Thus, despite the general dismissal of eliminative inference in the literature, it appears that this is in fact an inferential strategy that participants may form and use when responding in these tasks. Even so, it is clear that this inference is either used by only a very small minority of participants, or is one that is not readily articulated (i.e., it is quite possible that using an eliminative inference is more common than reporting it).

There were no responses expressing a clear novelty-matching strategy (that is, that unusual or novel trial types somehow go with the rarer outcome), or a more general preference for a more novel outcome. Responses that mentioned novelty did so only in reference to the novelty of the rare predictor. That is, the rarity of cue C meant that its *associated* outcome was more likely. The novelty of cue C during training appears to lead to a strong link between C and the rare outcome, which could indicate either an inferential or associative process. This is similar to the responses given by three participants classified as an asymmetric representation, in which C indicated the rare disease more than B indicated the common disease. Asymmetric representation will be discussed in greater depth in the following section of this paper.

While it is difficult to make strong conclusions given the small sample of rare choice explanations, these results do suggest that: (1) eliminative inference is a plausible, but not commonly reported, inference, (2) the novelty of C is important for the effect, and (3) there is little evidence for a unified inferential process resulting in rare outcome choice.

The evidence for reasoning in the effect is not clear cut, and may point to the contribution of multiple processes. Several researchers, including those who have proposed associative and attentional accounts of the inverse base-rate effect, have acknowledged a potential role for higher-order reasoning in the effect (Johansen et al., 2007; Kruschke, 2003; Kruschke, 2005; Winman, Wennerholm, & Juslin, 2003). Dual-process accounts, which assume associative learning processes may be influenced by, or interact with, higher-order inferential processes, still form the most complete explanations for other learning and cue competition effects (McLaren et al., 2014; Thorwart & Livesey, 2016). The assumption is that associations and inferences describe general mechanisms jointly governing human learning and behaviour in most situations, and thus it seems unlikely that the inverse base-rate effect should be any different in this regard.

The mixed findings regarding the contribution of higher-order processes highlight the need for further research into individual differences in the effect, to determine which factors predict a predominance of inferential or associative processes, and whether the tendency to rely on either process results in differences in choice. The tendency to choose the rare outcome might be associated with different learning orientations or cognitive abilities. For example, prior research has shown associations between the use of higher-order processes and cognitive reflection (Don, Goldwater, Otto, & Livesey, 2015; Don et al., 2016; Livesey, Lee, & Shone, 2013), and interactions between learning orientation and task conditions (Goldwater et al., 2018). If there are individual differences in the proclivity for rule use, then manipulations that increase the obviousness of rules or encourage their use should affect individuals differently (e.g., Don, Goldwater, Greenaway, Hutchings, & Livesey, 2020; Goldwater et al., 2018). These kinds of manipulations may help us draw inferences above and beyond the simple correlation found by Winman et al. (2005).

Asymmetric representation and the role of similarity

Kruschke (2001a) suggests that attention shifts during learning result in an asymmetric representation of cue-outcome relationships. Cue A receives more attention on AB trials than it does on AC trials, such that both A and B predict O1, but only C predicts O2. In category learning terms, the O1 category is represented by both A and B, whereas the O2 category is represented by C alone. On BC transfer trials, the absence of cue A means that BC is more similar to the representation for O2 than the representation for O1 (e.g., Nosofsky, 1984, 1986). Johansen et al. (2007) trained participants in a "disjoint cue structure" that mirrors this asymmetric structure, where AB-O1 was presented three times as often as C-O2. This resulted in greater O2 choice on BC trials. Given that participants did not show a choice bias when given a summary of the base-rates in an explicit decision-making task, they concluded that both asymmetric association and base-rate neglect were individually necessary and jointly sufficient to produce the inverse base-rate effect.

In our questionnaire study, three participants provided explanations classified as asymmetric representation, suggesting that C indicates the rare outcome more than B indicates the common outcome (although a small sample, these participants also showed a strong rare bias). This is more difficult to classify as a purely inferential strategy. Asymmetric representation of the information presented during training may well be a feature of the decisions that lead to the inverse base-rate effect, no matter what cognitive processes are responsible for those decisions. Associative networks that account for the effect (e.g., Kruschke's EXIT model) produce asymmetries because both the attention to cues and the associations with O1 are distributed across A and B on common AB-O1 training trials, whereas attention and O2 associations heavily favour cue C on the rare AC-O2 trials. However, reasoning processes may also be influenced by biased attention and asymmetric representation. Take the process of elimination, for example, where it is assumed that participants reject the common outcome because they notice that BC trials are different from well-learned instances of AB-O1. Asymmetric representation may contribute to people regarding AB and BC as being less similar than AC and BC. If it is assumed that C is represented in a way that makes it individuated and distinct, then AC and BC at least share a relevant cue in common. In contrast, if A and B are only retrieved as part of an integrated representation of AB-O1, participants may not make the same link between AB and BC, and may regard them as being distinctly different.[4]

[4] While this is possible, O1 responses to AX trials suggest this might not be the case (Don & Livesey, 2017).

Both inferential reasoning and associative memory may be affected by changes in the psychological similarity between trained and test trials. Associative accounts make the limiting assumption that generalization from past to present conditions occurs on the basis of similarity, and particularly on the basis of the presence of common features that have been associated with outcomes in the past. Inferential reasoning could also occur on the basis of feature similarity – indeed, associative generalization could simply be viewed as a form of inductive reasoning – but it may also occur on the basis of other factors, including *dissimilarity*. Elimination is an example of this; it is assumed that an individual will not choose the common outcome if they notice that the conflicting BC trial is substantially different from the AB-O1 trials they have experienced previously. Differences between similarity and dissimilarity processes have been captured formally in an extension of the generalized context model, known as the dissGCM (Stewart & Morin, 2007). The dissGCM incorporates standard computations of similarity to memories of learned exemplars (Nosofsky, 1986), with calculations of dissimilarity to those exemplars, with both contributing to decisions about category membership. O'Bryan et al. (O'Bryan et al., 2018) used this model to find independent neural correlates of similarity and dissimilarity processes while participants performed an inverse base-rate task. While it remains open to debate whether similarity and dissimilarity processes reflect the operation of qualitatively distinct psychological operations (like reasoning and associative memory), they do at least appear to engage different neural circuits.

Measurement of the inverse base-rate effect

There is documented variation in the strength of inverse base-rate effects across studies, and some researchers have argued that the effect may be limited to a particular set of conditions (Winman et al., 2005). To make this conclusion with confidence, we would need to know that the methodological parameters were optimised to find the effect, yet these parameters have not been systematically explored in a satisfactory way. The following section therefore reviews the parameters that have been varied in inverse base-rate research.

Base-rates

In a typical inverse base-rate effect task, AB-O1 trials are presented more frequently than AC-O2 trials. This basic design is often repeated multiple times with different cues and outcomes, including two (e.g., Bohil et al., 2005; Kruschke, 1996; Kalish, 2001; Lamberts & Kent; Wood, 2009; Wood & Blair, 2011), three (e.g., Medin & Edelson, 1988; Winman et al., 2005) or four (e.g., Don & Livesey, 2017; Don et al., 2019a; Medin & Bettger, 1991; Juslin et al., 2001) repetitions.

Although most tasks have used a 3:1 base-rate of common to rare trials in training, the effect has also been demonstrated with a more extreme 7:1 base-rate (Juslin et al., 2001; Lamberts & Kent, 2007; Shanks, 1992). Some evidence suggests that the 7:1 base-rate leads to a stronger effect than a 3:1 base-rate when trained within-subjects (Shanks, 1992). However, the effect is also observed with a weaker 2:1 base-rate (Wills et al., 2014). The effect also persists when the relative base-rates of common and rare trials are changed at different stages throughout training, with a choice bias for the early rare outcome on BC trials (Kruschke, 2009; Medin & Bettger, 1991). Based on these findings, Kruschke (1996, 2009) has suggested that learning the common contingencies before the rare contingencies is critical for the effect. Indeed, training AB-O1 trials prior to the introduction of AC-O2 trials also leads to a preference for O2 on BC trials in the *highlighting* effect (Kruschke, 1996).

Task scenarios

Studies of the inverse base-rate effect have traditionally used a medical diagnosis task, where participants assume the role of a doctor learning to diagnose fictitious diseases from experience with several patients (e.g., Johansen et al., 2007; Juslin et al., 2001; Kruschke, 1996; Medin & Bettger, 1991; Medin & Edelson, 1988). Nevertheless, the effect also appears to be robust under a variety of task scenarios and stimuli, including random word associations (Dennis & Kruschke, 1998; Kruschke, 2005), abstract shapes and coloured squares (Fagot et al., 1998), line stimuli (Kalish & Kruschke, 2000; Johansen et al., 2010), features of cell images and viruses (Lamberts & Kent, 2007), abstract shapes representing "cell bodies" and diseases (Wills et al., 2014), foods and allergic reactions (Don et al., 2019a), graphs of "blood proteins" and native Australian animal species (Kalish, 2001), as well as personality traits and group membership (Sherman et al., 2009). Other cue competition effects that have been linked to the inverse base-rate effect (e.g., blocking; Kruschke et al., 2005) seem to be affected by the explicit causal relationship implied by the task scenario (e.g., Blanco, Baeyens, & Beckers, 2014; Don & Livesey, 2018; Luque, Cobos, & López, 2008; Waldmann, 2000, 2001; Waldmann & Holyoak, 1992). The robustness of the effect across various task scenarios suggests this may not be the case for the inverse base-rate effect, although this has yet to be directly tested.

Test measures

With the exception of one study that used a cued-recall paradigm (Dennis & Kruschke, 1998), the inverse base-rate effect has been measured using a discrete choice between outcomes, where participants are asked to select the most likely outcome.[5] That is, participants are asked to choose which outcome they think is most likely, given the presented cues. Some studies have additionally included confidence ratings with transfer trial responses (Don & Livesey, 2017; Don et al., 2019a), which often reveal relatively high levels of confidence in choices on conflicting trials.

Given that several other cue competition effects, such as blocking, have been shown across different kinds of test measures (see Don & Livesey, 2018; Jones, Zaksaite, & Mitchell, 2019; Livesey, Greenaway, Schubert & Thorwart, 2019; Luque, Vadillo, Gutiérrez-Cobo, & Le Pelley, 2016; Mitchell, Lovibond, & Gan, 2005; Mitchell, Lovibond, Minard, & Lavis, 2006, for examples using predictive ratings, causal ratings, and discrete outcome choice), it is noteworthy that the inverse base-rate effect has almost exclusively been measured using discrete outcome choice. If the inverse base-rate effect is a result of similar mechanisms to other cue competition effects, we might expect it to occur under the wide variety of test conditions used in other cue competition effects. However, currently, there is a lack of studies demonstrating the effect (or its absence) using other test measures, even those that are commonly used in contingency learning experiments, such as continuous predictive ratings.

Statistical analysis

There is little consensus about the most appropriate way to statistically analyse the inverse base-rate effect. Many have adopted the approach of using simple paired t-tests comparing choice of common and rare outcomes directly, even though the underlying distributions of choice probability may deviate from normality. Several others have used chi-square tests, yet typically each participant contributes more than one response for each trial type (there are several variants of BC trials), which constitutes a violation of the assumption of independence required for chi-square tests. Dennis and Kruschke (1998) addressed this issue stating the chi-square values they obtained were large enough that conservative adjustments would still lead to the same conclusions, however, this may not always be the case with weaker effects.

Consistent with the analyses just described, most studies assume (either tacitly or explicitly) that a meaningful inverse base-rate effect is identified relative to a point of impartiality, where participants favour neither the rare nor the common outcome over the other, and therefore an inverse base-rate effect occurs when choice of the rare outcome is above chance. Some studies have instead compared choice on BC

[5] An unpublished study by Wedell and Kruschke (2001, as cited by Kruschke, 2009) measured likeability ratings in a task where participants used personality traits to predict group membership.

trials to choice on A trials (Lamberts & Kent, 2007; O'Bryan et al., 2018). The imperfect predictor may provide an appropriate comparison for rational base-rate use, as the frequency with which A is paired with each outcome is equivalent to the overall base-rate of each outcome. However, while one rational response to BC trials would be to predict the common outcome, another rationally justifiable response would be to show no choice bias, as the probability of O2 given C is equivalent to the probability of O1 given B. Thus, this statistical approach may lend itself to showing a significant inverse base-rate effect when choice of the common and rare outcome on BC trials do not differ, but there is

Table 4 Summary of empirical findings, strength of evidence, and consistence with theoretical accounts

Empirical findings			Consistency with theoretical accounts			
Observation	Strength of evidence	Notes	Attention + Prediction error		Elimination inference (ELMO)	Novelty selection/ Novelty matching
			Simple (Le Pelley et al., 2016)	Complex (EXIT)		
Rare bias on BC trials	Strong[a]	*The inverse base-rate effect, observed many times*	Con	Con	Con	Con
Common bias on A trials	Strong[a]	*Almost always accompanies the inverse base-rate effect*	Con	Con	Con	Con
Generality across task scenarios	Strong[a]	*Widely replicated in different tasks*	Con	Con	n/a	n/a
Necessity of shared cue	Moderate[b]	*Consistent result in three studies*	Con	Con	Inc	Inc
Generality across biased base-rates	Moderate[b]	*Replicated in base rates ranging from 2:1 to 7:1*	Con	Con	Con	Con
Attention bias C > B	Moderate[b]	*Tested once in eye-gaze and associability (two similar effects in highlighting)*	Con	Con	n/a	Con
Global outcome base-rate effect	Moderate[b]	*Consistent result in three studies*	Inc*	Con**	Con	Con
Higher confidence on conflicting than imperfect or novel trials	Moderate[b]	*Consistent pattern in two studies, statistically analysed in one.*	Con	Con	Inc	Inc
Common bias on ABC trials	Moderate[c]	*Often tested, sometimes shows rare bias*	Con	Con	Con	Inc
Accuracy: B > C	Moderate[c]	*Evident in many (not all) studies but rarely specifically tested*	Inc*	Con**	Con	n/a
Rare bias on novel cue trials	Weak[e]	*Two studies, conflicting results*	Inc	Inc	Con	Con
Common bias on novel AX trials	Weak[d]	*One study only, replicated in three experiments*	Con	Con	Inc	Inc
(lack of) effect of cognitive load	Weak[d]	*One study only, weak rare bias*	Con+	Con+	Inc	Inc
(lack of) effect when cue frequency is matched and outcome base-rate is present	Weak[f]	*One study only*	Con	Con	n/a	Inc
Common predictor processing at test	Weak[f]	*One study only, replicated within the study, weak rare bias*	Inc	Inc	Con	Inc
Correlation with rule use in another task	Weak[f]	*One study only*	Inc+	Inc+	Con	Con
(lack of) effect in children	Weak[f]	*One study only, weak age difference (conflicting result in highlighting)*	Inc+	Inc+	Con	Con
(lack of) effect in other animals	Weak[f]	*One study in highlighting only with N=2*	Inc+	Inc+	Con	Con
Discrete cues not required	Weak[f]	*One study only*	n/a	Con	Inc	Inc
Deterministic relationships required	Weak[f]	*One study only, unusual design*	Inc	Inc	Con	n/a

Note: Inc = Inconsistent; Con = Consistent. Strength of evidence is classified on the following bases: (Strong) [a] Widely replicated across many studies; (Moderate) [b] Consistent effects / large effect sizes across relatively few studies, [c] Fairly consistent (with some exceptions) over many studies; (Weak) [d] Replicated multiple times within one study only; [e] Inconsistent results over relatively few studies, [f] Only a single study involving the inverse base-rate effect. * Could be consistent if we assume context associations are involved. ** Likely requires learning about context due to the presence of a bias cue in the EXIT model. + This assumes that associative processes are bottom-up and automatic in nature and thus should have primitive origins and no relationship to higher-order cognition or working memory. However, the validity of these assumptions is highly debatable and there is no reason to assume that selective attention in particular should have these properties. These manipulations can only really provide evidence against the inferential accounts that assume memory dependence

Table 5 Outstanding questions and possible avenues for further research

Research question	Theoretical interpretation
How are conflicting BC trials processed at test?	Evaluating whether there is greater attention to the common or rare predictor at test will help determine whether the effect is a result of facilitating rare responses or inhibiting common responses
What is the role of context associations in the inverse base-rate effect?	Strong associations between the context and the common outcome appear to contribute to attention biases favouring rare predictors, but there is still mixed evidence regarding the effect of context on test trials and counterintuitive B>C effects
Are attention biases to rare predictors during training driven by novelty or prediction error (or both)?	Several results indicate that greater attention is paid to the rare predictor. This may be because the predictor itself is surprising due to its novelty, or because its associated outcome is surprising (i.e., the predictor is accompanied by larger prediction error). EXIT assumes the latter; however, novelty and prediction error are confounded in these tasks
How well does the EXIT model account for other attention-based learning effects?	If the inverse base-rate effect is useful for theories of attention-based learning, the model that best accounts for it should also account for other attention-based learning effects
How does attention during feedback differ from attention prior to making a prediction?	Research in the inverse base-rate effect indicates different attention biases prior to making a prediction and during feedback. A question remains as to whether this occurs in other attention-based learning effects
Does the inverse base-rate effect occur with probabilistic contingencies?	This will assess the generalisability of the effect. If the effect does not occur with probabilistic contingencies, this would be inconsistent with connectionist accounts
Which individual differences predict the inverse base-rate effect?	Correlations with executive cognitive processes, including abstract reasoning and cognitive control of attention will help determine whether rule-based (or other executive) processes may be involved in the effect. Manipulations encouraging rule use could serve a similar aim
Does the inverse base-rate effect occur in animals?	If the effect occurs in animals in simple conditioning paradigms, this would provide good support for a generalisable competitive learning effect
Does the inverse base-rate effect occur in children?	Further research is needed to test whether the absence of the effect in children is consistent. If the effect does occur in children, this might call into question whether high-level reasoning processes are necessary
Does dissimilarity-based elimination play a more general role in learning from experience?	The eliminative inference is not well supported as an explanation of the inverse base-rate effect, however perhaps a more general dissimilarity decision mechanism is important here and in other decisions based on prior experience. More evidence and theory development is needed
Does the inverse base-rate effect occur in different types of test judgements?	Other cue competition effects tend to be demonstrated in different types of judgements, e.g. learned predictiveness and blocking effects

a common bias on A trials. Comparisons of choice on BC trials against A therefore provide a good test of base-rate neglect, while comparisons against chance provide a stronger test of an *inverse* base-rate effect.

Conclusions and future directions for the inverse base-rate effect

Table 4 summarises the key empirical results related to the inverse base-rate effect that we have addressed, the strength of evidence for each of these results, and their consistency with different theoretical accounts. Table 5 summarises key questions where further evidence is needed. On the whole, the effect seems unlikely to be a result of a highly specific inference such as eliminative inference. The data presented here show that eliminative inference is an occasionally reported strategy in this task, although it is unlikely to account for the range of phenomena associated with the inverse base-rate effect. It also suggests the inverse base-rate effect can be produced in the absence of this strategy. While we have attempted to test alternative inferences that may result in the inverse base-rate, for example, novelty selection or novelty matching (Don & Livesey, 2017), this inference suffers similar limitations in terms of its inability to account for the available data. Our sample of participants who described their inferential choices in a post-experiment questionnaire offered no indication of any other inference that they could articulate or consistently use. In addition, participants generally tend to be more confident in their choices on conflicting BC trials than on imperfect and novel transfer trials (Don & Livesey, 2017; Don et al., 2019a), which is inconsistent with the idea that participants are simply making odd choices as a result of uncertainty about these trials. For now, far and away the most defensible conclusion is that the inverse base-rate effect is not the consequence of a single specific and idiosyncratic inference like elimination, nor is it obviously the consequence of a collection of well-articulated inferences that each possess

similar specificity (and thus would not be found elsewhere in learning and decision making). While the eliminative inference as it has been applied formally to the inverse base-rate effect is highly specific (and proponents of its use have argued that the effect itself might be highly idiosyncratic), a general form of this cognitive mechanism is still worth taking seriously. That is, decisions and judgments underpinned by dissimilarity-based elimination may be more widely used in human cognition than our analysis of the inverse base-rate effect suggests. It remains a challenge for future research to find stronger evidence of such processes, particularly in the type of learning and decision tasks that we are concerned with here. Assuming such evidence can be found, determining how such decisions might contribute to the inverse base-rate effect is still a problem that needs to be resolved. In our opinion, there are too many inconsistencies between the eliminative inference and empirical evidence for this to be considered a viable option.

Currently, the empirical results that are inconsistent with attention-based models of the effect either have weak or conflicting evidence, or can be accounted for if we make additional, reasonable assumptions, such as learning about context associations. The more general evidence that the inverse base-rate effect requires executive function or higher-order cognition (e.g., correlation with rule discovery and developmental trajectory) only constitutes evidence against attention-based prediction error models if one makes the assumption that selective attention in learning is independent of cognitive resources. We suggest that there is little basis for this assumption; selective attention processes may well rely on executive functions even if basic association formation does not. There is some evidence for automaticity in the attentional changes that are driven by predictive learning but there is also evidence of cognitive control (e.g., Mitchell, Griffiths, Seetoo & Lovibond, 2012).

To know whether the inverse base-rate effect reflects broad cognitive mechanisms, it is critical to determine the generality of the effect. While the effect occurs across different task scenarios and conditions, some conditions have remained consistent across all studies and may be unlikely to arise in the real world. For instance, almost all studies have used perfect contingencies between predictive cues and outcomes. One study has shown that the effect disappears when using continuous overlapping cue magnitude distributions. The interpretation they offer is that this renders the cue-outcome relationship probabilistic rather than deterministic. Thus, testing whether preferences for rare outcomes still occur with probabilistic cue-outcome relationships is an important aim for future research. In addition, the effect has typically been assessed using a binary choice test. It is therefore important to determine whether the effect occurs in different types of test judgements, for example, predictive or causal ratings.

While there is consistent evidence of attention biases during training, there are currently mixed results regarding how conflicting BC trials are processed at test. To better determine the nature of decision processes happening on these trials, future research requires complementary evidence from eye-tracking, associability measures, and imaging to determine how participants make decisions on these trials, for instance, whether the rare-outcome choices are a result of facilitating rare responses by attending to the rare predictor, or inhibiting common responses after attending to the common predictor.

Further work is also needed to examine the role of context associations in driving outcome predictions, since learning about context associations may be necessary to explain some properties of the effect, and the evidence for such learning is still minimal. While the EXIT model is generally successful in accounting for many of the phenomena accompanying the inverse base-rate effect, the model is complex, potentially to the detriment of understanding how individual theoretical mechanisms are useful in generating the effect. If the inverse base-rate effect is important for the development of theories of attention-based learning, then the capabilities of the EXIT model also require broader testing, in order to determine whether it serves as an effective general-purpose model of attention. Recent work by Paskewitz and Jones (2020) has begun to address this. Future work is needed to test how well this model or a reduced version of it can account for other attention-based effects, for example, effects driven by absolute or relative predictiveness, or uncertainty. Further, it will be important to establish whether attention biases to rare predictors are based on prediction error specifically, or simply cue novelty. The EXIT model predicts that attention is driven by prediction error, but this is confounded with cue novelty in the task.

Conclusions about the triviality of the effect (e.g., Winman et al., 2005) are currently based on relatively weak evidence. In some cases, there is simply scarce evidence on which to make conclusions. For instance, further research in children and animals is required to ascertain whether the absence of the inverse base-rate effect in these populations is consistent. More generally, the link between the inverse base-rate effect and executive cognitive processes needs to be explored further, including determining the role played by abstract reasoning on the one hand and cognitive control of attention on the other. For instance, the correlations with rule-based processes require greater examination, and could be better assessed by task manipulations that highlight rules or encourage their use.

Finally, inconsistencies in findings may be reconciled by appealing to individual differences. There is need for greater assessment of whether differences in learning, cognitive ability, or cognitive strategies can predict the inverse base-rate effect, in order to fully understand the processes involved in the effect, and to identify which individuals may be most susceptible to this misuse of base-rate information.

Misuse of base-rate information in learning from experience

The inverse base-rate effect reveals conditions in which people consistently misuse base-rate information when learning from experience. The base-rates of events provide an important source of information when making decisions based on ambiguous information. Yet, people are often poor at making rational choices when they involve a consideration of base-rates. Despite previous suggestions that trial-by-trial learning may encourage rational use of base-rates (e.g., Christensen-Szalanski & Beach, 1982; Butt, 1988; Christensen-Szalanski & Bushyhead, 1981; Manis et al., 1980), this is clearly not the case in all situations.

On the whole, the evidence currently favours the inverse base-rate effect being driven by general learning and memory processes coupled with changes in attention, and thus should not be dismissed by theorists interested in these processes. The inverse base-rate effect is not the only learning phenomenon in which biased judgements arise from the erroneous use of base-rate information. Base-rate neglect occurs in trial-by-trial learning in a number of experimental contexts. For example, the formation of illusory causation between a cue and an outcome when there is zero contingency (i.e., when the probability of the outcome is the same when the cue is present and when the cue is absent) is assumed to be the result of failing to adequately consider the base-rates of the outcome when the cue is absent (e.g., Matute et al., 2015). This illusory causation is also greater when the outcome is experienced frequently (Blanco, Matute, & Vadillo, 2013). It is possible that these effects are stronger in trial-by-trial learning conditions than when participants are made explicitly aware of the base-rates. Indeed, when participants are pre-trained to expect a high base-rate of the outcome, illusory causation is reduced (Blanco & Matute, 2019). This may have a similar effect to providing participants with a summary of the base-rates, as in Johansen et al. (2007). Reward frequency also appears to have stronger influences on choice behaviour than reward probability in reinforcement learning tasks (Don, Otto, Cornwall, Davis, & Worthy, 2019b). It remains to be seen whether these errors of judgements are the consequence of the same cognitive mechanisms (see Kutzner & Fiedler, 2015; Sherman et al., 2009), but at the very least, they seem to arise in similar learning contexts.

Curiously, the conditions in which the inverse base-rate effect is most reliable are also the conditions in which participants display the greatest learning of base-rates in other ways. For instance, rare biases on conflicting trials are stronger when there are greater common responses to the imperfect predictor (Shanks, 1992). Indeed, conditions that reduce the rare bias on conflicting trials also reduce common responses on other trials (e.g., Don & Livesey, Experiments 2 and 3). The effect therefore does not appear to be a result of failing to learn base-rate information, but is instead a consequence of learning the base-rates well. Perhaps above all else, the effect highlights that efficient learning mechanisms can lead to faulty decisions. Studying these seemingly puzzling decisions provides an opportunity to understand the mechanisms by which we learn and use frequency information in day-to-day life.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-020-01870-0.

References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211–33.

Bar-Hillel, M., & Fischhoff, B. (1981). When do base rates affect predictions? *Journal of Personality and Social Psychology, 41*, 671–80.

Binder, A., & Estes, W. K. (1966). Transfer of response in visual recognition situations as a function of frequency variables. *Psychological Monographs, 80* (23, Whole No. 631).

Blanco, F., Baeyens, F., & Beckers, T. (2014). Blocking in human causal learning is affected by outcome assumptions manipulated through causal structure. *Learning & Behavior*, *42*, 185–199.

Blanco, F., Matute, H., & Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior*, *41*, 333–340.

Blanco, F., & Matute, H. (2019). Base-rate expectations modulate the causal illusion. PloS one, 14, e0212615.

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

Burling, J. M., & Yoshida, H. (2016). Highlighting in early childhood: Learning biases through attentional shifting. *Cognitive Science*, *41*, 96–119.

Butt, J. (1988). Frequency judgments in an auditing-related task. *Journal of Accounting Research, 26*, 315–30.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*, 999–1001.

Christensen-Szalanski, J. J. J. & Beach, L. R. (1982) Experience and the base-rate fallacy. *Organization Behavior and Human Performance, 29*, 270-78.

Christensen-Szalanski, J. J. J. & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 928–35.

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

Denton, S. E., & Kruschke, J. K. (2006). Attention and salience in associative blocking. *Learning & Behavior, 34*, 285–304.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology, 37*, 397–416.

Dickinson, A., Shanks, D. R., & Evenden, J. L. (1984). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A*, 29-50.

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. (2018). Is the blocking effect sensitive to causal model? It depends how you ask. *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp.1633–1638). Madison: Cognitive Science Society

Don, H. J. & Livesey, E. J. (2021). Attention biases in the inverse base-rate effect persist into new learning. *The Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021820985522.

Don, H. J., Goldwater, M. B., Otto, A. R., & Livesey, E. J. (2015). Connecting ruleabstraction and model-based choice across disparate learning tasks. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 590–595). Pasadena, California: Cognitive Science Society.

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

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

Don, H. J., Otto, A. R., Cornwall, A. C., Davis, T., & Worthy, D. A. (2019b). Learning reward frequency over reward probability: A tale of two learning rules. *Cognition*, *193*, 104042.

Don, H. J., Goldwater, M. B., Greenaway, J. K., Hutchings, R., & Livesey, E. J. (2020). Relational rule discovery in complex discrimination learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 1807–1827.

Fagot, J., Kruschke, J. K., Dépy, D., & Vauclair, J. (1998). Associative learning in baboons (Papio papio) and humans (Homo sapiens): Species differences in learned attention to visual features. *Animal Cognition, 1*, 123–133.

Gluck, M. A. (1992). Stimulus sampling and distributed representation in adaptive network theories of learning. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes, Vol. 1. From learning theory to connectionist theory; Vol. 2. From learning processes to cognitive processes* (pp. 169–199). Hillsdale: Lawrence Erlbaum Associates.

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

Goldwater, M. B., Don, H. J., Krusche, M., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General.*

Granger, K. T., Moran, P. M., Buckley, M. G., & Haselgrove, M. (2016). Enhanced latent inhibition in high schizotypy individuals. *Personality and Individual differences, 91*, 31-39.

Hamilton, D. L. (Ed.) (1981). *Cognitive processes in stereotyping and intergroup behavior*. Hillsdale: Erlbaum.

Haselgrove, M., Le Pelley, M. E., Singh, N. K., Teow, H. Q., Morris, R. W., Green, M. J., ... & Killcross, S. (2016). Disrupted attentional learning in high schizotypy: Evidence of aberrant salience. *British journal of psychology, 107*, 601–624.

Inkster, A., Milton, F., Edmunds, C. E. R., Benattayallah, A., & Wills, A. (2019a) *Neural Correlates of the Inverse Base Rate Effect*. https://doi.org/10.31234/osf.io/muqrh

Inkster, A., Mitchell, C., Schlegelmilch, R., & Wills, A. (2019b). Effect of a context shift on the inverse base rate effect. https://doi.org/10.31234/osf.io/rpb7x

Johansen, M. K., Fouquet, N., & Shanks, D. R. (2007). Paradoxical effects of base rates and representation in category learning. *Memory & Cognition, 35*, 1365–1379.

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., Zaksaite, T., & Mitchell, C. J. (2019). Uncertainty and blocking in human causal learning. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*, 111.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*, 237–251.

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.

Kamin, L.J. (1969). Selective association and conditioning. In N.J. Mackintosh & W.K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 42–64). Halifax: Dalhousie University Press.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioural and Brain Sciences, 19*, 1–53.

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

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

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

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

Kruschke, J. K. (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.

Kutzner, F. L., & Fiedler, K. (2015). No correlation, no evidence for attention shift in category learning: Different mechanisms behind illusory correlations and the inverse base-rate effect. *Journal of Experimental Psychology: General, 144*, 58.

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

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

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (2018). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 1955–1970.

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

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

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

Le Pelley, M. E., Schmidt-Hansen, M., Harris, N. J., Lunter, C. M., & Morris, C. S. (2010). Disentangling the attentional deficit in schizophrenia: Pointers from schizotypy. *Psychiatry Research, 176*, 143–149.

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & cognition*, *43*, 283–297.

Livesey, E. J., Greenaway, J., Schubert, S., & Thorwart, A. (2019). Testing the deductive inferential account of blocking in causal learning. *Memory & Cognition, 47*, 1120–1132.

Livesey, E., Lee, J., Shone, L. (2013). The relationship between blocking and inference in causal learning. *35th Annual Meeting of the Cognitive Science Society (COGSCI 2013)*, Austin: Cognitive Science Society.

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

Luque, D., Vadillo, M. A., Gutiérrez-Cobo, M. J., & Le Pelley, M. E. (2016). The blocking effect in associative learning involves learned biases in rapid attentional capture. *The Quarterly Journal of Experimental Psychology*, 1–26.

Luque, D., Cobos, P. L., & López, F. J. (2008). Interference between cues requires a causal scenario: Favorable evidence for causal reasoning models in learning processes. *Learning and Motivation*, *39*(3), 196–208.

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

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology*, *38*(2), 231248.

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.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in psychology*, *6,* 888.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*, 668–693.

McLaren, I. P., Forrest, C. L. D., McLaren, R. P., Jones, F. W., Aitken, M. R. F., & Mackintosh, N. J. (2014). Associations and propositions: The case for a dualprocess account of learning in humans. *Neurobiology of learning and memory*, *108*, 185–195.

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.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*, 207.

Mitchell, C. J., & Le Pelley, M. E. (Eds.). (2010). *Attention and associative learning: From brain to behaviour*. Oxford University Press, USA.

Mitchell, C. J., Lovibond, P. F., & Gan, C. Y. (2005). A dissociation between causal judgment and outcome recall. *Psychonomic bulletin & review*, *12*, 950–954.

Mitchell, C. J., Lovibond, P. F., Minard, E., & Lavis, Y. (2006). Forward blocking in human learning sometimes reflects the failure to encode a cue–outcome relationship. *The Quarterly Journal of Experimental Psychology*, *59*, 830-844.

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.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*, 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*, 39.

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.

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.

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: A review and a possible integration. *Attention and associative learning: From brain to behaviour*, 11–39.

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

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.). *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms (No. VG-1196-G-8). Cornell Aeronautical Lab Inc Buffalo NY.

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

Shanks, D. R., & Darby, R. J. (1998). Feature-and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 405–415.

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.

Stewart, N., & Morin, C. (2007). Dissimilarity is used as evidence of category membership in multidimensional perceptual categorization: A test of the similarity–dissimilarity generalized context model. *Quarterly Journal of Experimental Psychology, 60*, 1337–1346.

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

Thorwart, A., & Livesey, E. J. (2016). Three ways that non-associative knowledge may affect associative learning processes. *Frontiers in psychology*, *7*, 2024.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*, 207–232.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

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

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 53–76.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8*, 600–608.

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

Widrow, B., & Hoff, M.E. (1960). Adaptive switching circuits. 1960 WESCON Convention Record Part IV, 96–104.

Wills, A. J., Barrasin, T. J, & McLaren, I. P. L. (2011a). Working memory capacity and generalization in predictive learning. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3205–3210). Austin: Cognitive Science Society.

Wills, A. J., Graham, S., Koh, Z., McLaren, I. P., & Rolland, M. D. (2011b). Effects of concurrent load on feature-and rule-based generalization in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(3), 308–316.

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

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

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

Wood, M. J. (2009). *Categorization of partially occluded visual stimuli: bridging the gap between completion and classification* (Doctoral dissertation, Dept. of Psychology-Simon Fraser University).

Wood, M. J., & Blair, M. R. (2011). Informed inferences of unknown feature values in categorization. *Memory & cognition*, *39*, 666–674.