Errorless irrationality: removing error-driven components from the inverse base-rate effect paradigm

Anonymous CogSci submission

Abstract

Include no author information in the initial submission, to facilitate blind review.

Keywords: irrationality; prediction error; inverse base-rate effect; categorization; contingency learning

Introduction

The *inverse base-rate effect* (IBRE, Medin & Edelson, 1988) is an irrational tendency in humans to overweigh rare events when faced with ambiguity. In a traditional design, people learn to categorise two overlapping sets of features under two distinct labels. These sets share a single feature, A, and possess a unique feature, B and C, predictive of their respective category label. During learning, these sets of features occur at different frequencies. The features under the common label usually occur three times as much as features under the rare label (Kruschke, 1996). Following training, people categorise features presented by themselves and un unique combinations. People optimally label uniquely predictive features, B and C, presented individually with their respective common and rare labels. Responses on the shared feature A also show base-rate following. But when uniquely predictive features are paired, B and C, people tend to respond with the rare category label. According to Classical Probability Theory, the rational response is to attribute the common label to this ambiguous compound, because it is the most frequently occurring label. This rare bias on ambiguous combinations of BC has been observed across a different variety of experimental manipulations (Kalish, 2001; Don & Livesey, 2017, 2021; Inkster, Mitchell, Schlegelmilch, & Wills, 2022; Wills, Lavric, Hemmings, & Surrey, 2014). For a more thorough introduction to this irrational bias, see a review by Don, Worthy, and Livesey (2021).

Assumptions of theories of the IBRE

The most prominent theories of the inverse base-rate effect involve an attentional mechanism that drives not only learning but responding as well. These theories are formal models. They are: a neural network with an exemplar-mediated attention to distinctive input, EXIT (Kruschke, 2001b), a three-layer neural network with competitive attentional gating, and a four-layer neural network with an additional rapid attentional shift (Paskewitz & Jones, 2020). All these explanations rely on a process that reallocates attention in response to prediction errors - they update attentional values according to gradient descent. Their explanation is simple. During learning, people learn to label the AB compound first. They are still learning to label the AC compound, so when they make an error, attention relocates towards the uniquely predictive feature C to reduce future errors. This results in C acquiring higher attentional salience than B. When the ambiguous BC compound is presented, C will dominate responding, resulting in an irrational tendency to respond with the rare label. According to these models, this irrationality results from an optimisation process that tries to reduce the errors people make. This process results in an asymmetric representation that can be summarised as AB belongs to common, $AB \rightarrow common$, and C belongs to rare, $C \rightarrow rare$ (Kruschke, 2001a).

Current Study

In this work, we intend to test this basic assumption of these theories. In the following two experiments, we will gradually remove components from the design traditionally associated with prediction error. Our overarching goal is to investigate whether we can still observe the IBRE, even if we experimentally remove a crucial assumption of already existing accounts. In our first attempt, building on the observational learning condition of Experiment 2 in Johansen, Fouquet, and Shanks (2007), we implement the canonical IBRE design with a caveat that category labels are presented in unison with features.

In our second attempt, we further remove the causal relationship between features and category labels. The goal was to remove any design component that might affect attentional allocation or the development of asymmetric representation in response to errors. Any presumption of a causal relationship can inadvertently relocate attention in line with the direction of causality between features and labels.

Related Work

To our knowledge, there is only one attempt to implement the IBRE procedure without explicit feedback. In their Experiment 2, Johansen et al. (2007) tried to observe the inverse base-rate effect both in a predictive-learning condition and in an observational learning condition. In the observational learning condition, category labels and features were presented at the same time to participants. Their design involved disjoint cues (where categories shared no features in common), while the canonical design depends on a shared feature during training. According to the theories of the IBRE, this shared feature facilitates attentional relocation. This attentional tuning in turn pushes responding toward the rare label. In contrast, Johansen et al. (2007) trained participants on $AB \rightarrow common$ and $C \rightarrow rare$. Their design was optimized to investigate the hypothesised asymmetric cognitive representation of the two categories. As a result, the only instance when they observed a rare bias was when the common feature, B ,presented in compound during training was paired

with a rare, *C*, feature presented by itself during training. This provided evidence for the asymmetric cognitive representation (Kruschke, 2001a) hypothesised to develop during training. A simple neural-network explanation can account for this by positing that the rare feature developed stronger connections with the category label. Compound features share the connection with the category label, so any update to these connections dissipates between them. There is no need to relocate attention to reduce errors, therefore attention will not bias responding toward the rare label. But stronger $C \rightarrow$ *rare* feature-label connections will result in a rare bias on BC compounds during the test phase. In order to investigate the role of prediction error in response to feedback, we need to tweak their observational learning condition to conform to a more canonical implementation of the procedure. This will involve reimplementing the shared feature. In terms of a clear observational-learning version of the IBRE, Johansen et al. (2007) included the result of a short pilot experiment in their Appendix. There are no details about the procedure of this experiment, so we cannot make direct comparisons. We will follow up on that line of investigation.

Our current study modified the canonical IBRE procedure in an attempt to remove error-driven components from the design - which includes a shared cue. This contrasts Experiment 2 in Johansen et al. (2007), who tried to remove the shared cue that pushed attention towards C during training and resulted in the asymetric representation. Here, we strictly focus on prediction error concerning behaviour.

The only two studies directly looking at error-driven processes in the IBRE are Inkster, Milton, Edmunds, Benattayallah, and Wills (2022) and Wills et al. (2014). Wills et al. (2014) observed posterior selection negativity and concurrent frontal positivity for C relative to B, which gave evidence for an error-driven selective attentional learning process. Inkster, Milton, et al. (2022) carried out a direct investigation into brain regions underlying error-driven learning in the IBRE. Their ROI analysis explicitly targeted areas that were hypothesised to be involved in the computation of prediction error. They showed that these areas exhibited greater activation during the test phase for C relative to B with a presence of a shared cue during training. Both Wills et al. (2014) and Inkster, Milton, et al. (2022) gave strong evidence in support of error-driven attentional learning accounts of the IBRE. In our study, we will look for the effect while trying to take out prediction error from the experimental design. In contrast, they looked for the neural substrates of error-driven accounts of the effect.

Experiment 1

Below, we detail our first attempt to test whether we could observe the rare response bias without an explicit error-driven psychological mechanism. The design component which is most likely to result in any error-driven tuning is feedback. To remove feedback, we will present category labels with their respective features. We retain the sequential property of the experiment, which means that participants learn about feature and category relationships on a trial-by-trial basis.

Experiment 1 is a conceptual replication of an experiment included in the Appendix of Johansen et al. (2007). The only information available is the list of test items (23), the doubled-up design (2 sets of categories and features), and the sample size. We substantially simplified our implementation by removing the doubled-up design and reducing the number of test items to 6.

Method

Participants Participants were undergraduate students who received course credit for their participation. We recruited 169 participants online through the SONA recruitment system.

Apparatus The experiment was programmed in JsPsych (De Leeuw, 2015) to be run in a web browser. Participants completed the experiment on their personal computers. The experiment did not allow the use of tablets and smartphones.

Stimuli Category labels corresponded with response keys and were called Disease **Z** and Disease **L**. Category features were symptoms: fever, headache, and rash. These physical features were randomly allocated to abstract features, A, B, and C at the beginning of each session. Features and labels appeared in full sentences, such as '*John has fever and rash, which belongs to disease Z*'. Names were randomly drawn from a pool of male and female first names. The list was compiled from an online repository of popular baby names[1]. We selected the 50 most popular male and female names from 2021. Disease names corresponded to response keys and were randomly allocated to either the common or rare category label at the beginning of each session.

Procedure Table 1 summarises the abstract design of the experiment. This design is the simplest implementation of the IBRE procedure to date. Participants completed two phases: a training and a test phase. In the training phase, they encountered descriptions of people, the symptoms they experienced, and their respective diseases. These descriptions appeared in the format of '*John has fever and rash, which belongs to disease Z*'. Participants studied these examples and when they were ready to move on, they pressed the spacebar. They needed to complete reading the description within 5 seconds. If the 5 seconds threshold has passed, a screen appeared with the message '*Please respond faster!*'. In each training block, participants encountered 6 common diseases (common category exemplars) and 2 rare diseases (rare category exemplars). After the second block of training, participants were given a choice. They could either move straight to the test phase or complete another training block. There were a maximum of 5 blocks they could complete.

In the test phase, participants judged individual symptoms and novel combinations of old symptoms, see Table 1. Symp-

[1]The list was taken and later curated from a GitHub repository.

toms appeared in a sentence, such as '*John has a fever.*', with a prompt asking participants to label what disease the person has, '*Does the patient has disease Z or disease L?*'. Participants had to respond by pressing either Z or L on the keyboard. They had 10 seconds to do so, otherwise, a '*Please respond faster!*' message appeared. After the button press, there was no feedback, Each unique test item and training item (occurring in the test phase) was repeated 20 times. The test phase, therefore, included 120 trials, which were broken down into 5 blocks of 24 trials.

Table 1: Abstract design of Experiment 1 including both test and training phases.

Training (Relative Frequencies)	**Test**	
$AB \rightarrow common_1$ (x 3)	A, B, C,	
$AC \rightarrow rare_1$ (x 1)	AB, AC, BC	x 20

Analysis In order to test the presence of the IBRE, we calculated a Bayes Factor for a one-sample design. We calculate the probability of responding with the rare label on the critical BC test item, $P(rare|BC)$, for each participant. Then we tested this distribution of probabilities against the null, $mu = 0.5$, which denoted random responding. If the Bayes Factor fell below 1/3, we concluded that participants' responses are not different from random responding. If the Bayes Factor fell above 3, we concluded that participants' responses reliably differ from null. If the mean probability of $P(rare|BC)$ is higher than 0.5, we conclude that we observed the IBRE. Values lower than 0.5 would indicate rational responding. We used the method implemented in the BayesFactor R package (Morey & Rouder, 2022).

Exclusion To match performance with the predictive learning implementations of the IBRE, we decided to exclude participants whose test performance on the training items fell below 0.75 accuracy. We arrived at this threshold by testing all different levels of accuracy by calculating the Bayes factors for binomial data. We used the method implemented in BayesFactor R package (Morey & Rouder, 2022). If the Bayes Factor fell above 3, we concluded that we have sufficient evidence to believe the participant learned the training items.

Results and Discussion

After exclusion, 125 participants made it into our main analysis. In a summary, the qualitative pattern in our results corresponds to the base result of the IBRE. Table 2 shows the group-level probabilities for each item. Predictive features and training items are classified into their respective category. Participants exhibited a reliable common preference for A, $M_A = 0.68$, 95% HDI $[0.63, 0.73]$, $\text{BF}_{10} = 2.45 \times 10^7$. People explicitly followed the base rate - responded rationally according to Probability Theory. On the contrary, participants showed a reliable rare preference for BC, $M_{BC} = 0.67$, 95% HDI $[0.62, 0.72]$, $\text{BF}_{10} = 1.11 \times 10^7$. This gives us a sufficient amount of evidence to conclude that we have observed the IBRE.

Table 2: Group-level mean probabilities for each stimulus presented during the test phase in Experiment 1 after exclusion.

	$P(common)$	$P(rare)$
A	0.69	0.31
AB	0.94	0.06
AC	0.08	0.92
B	0.94	0.06
BC	**0.33**	**0.67**
C	0.04	0.96

Here, we report a successful conceptual replication of an observational learning IBRE procedure reported by (Johansen et al., 2007). In the current experimental design, the IBRE emerged in the absence of an explicit prediction error that drives the development of attentional allocation. This prediction error also drives the development of an asymmetric cognitive representation. All current theories of the IBRE rely on the assumption that this irrational rare preference arises as a result of optimising accuracy during the training phase. In the absence of this explicit prediction error, current theories cannot deal with encoding this type of input representation.

One shortcoming of the current design is that participants can still make predictions about feature–category on a trial-by-trial basis. Given that the general assumption is that diseases cause symptoms, participants could likely assume a causal link between symptoms and diseases. This assumed causal relationship can encourage participants to make not an explicit (responding with the category label via the keyboard) but an implicit prediction. Informally, participants might think of a certain feature–label causal relationship while reading the sentences. People then resolve errors between the expected and the observed feature–label causality by allocating attention to rare features to distinguish diseases.

A simple solution is to remove any design component that makes it clear to participants what the category label is. In addition, stimuli need to be constructed in a way that reduces the chance of people assuming any causal relationship between its features.

Experiment 2

In this experiment, we implemented the IBRE in a way most similar to cued recall tasks. Previous category labels were treated as features. And features were selected to be solid black geometric shapes. The task asked participants to memorise the arrangement of these shapes. On each trial, we randomised the position of the geometric shapes in the arrange-

Figure 1: Simple geometric shapes used as stimuli in Experiment 2.

ment. This further minimised the chances of having any design component suggestive of which feature is the category label.

Method

Participants We recruited 65 undergraduate students who completed the experiment for partial course credit. Recruitment was done via the SONA recruitment system.

Stimuli Stimuli were common solid geometric shapes, shown in Figure 1. Common and rare category labels were turned into features X and Y respectively. Each shape was randomly allocated to one of the abstract features shown in Table 3.

Table 3: Abstract design of Experiment 2 including both test and training phases. X and Y are in place of the category labels. During the test phase, participants needed to select either X or Y to complete the features shown below.

Training (Relative Frequencies)	**Test**
ABX x 3	A, B, C,
ACY x 1	AB, AC, BC x 20

Procedure Table 3 depicts the abstract experiment design. Similar to the previous experiment, participants completed two phases: an encoding/training and a cued-recall/test phase. In the training/encoding phase, participants were repeatedly exposed to the exemplars and were asked to memorise the arrangement of geometric shapes. Compared to Experiment 1, exemplars were composed of three geometric shapes. On each trial, geometric shapes appeared in random order so the position of features on the screen was completely counterbalanced. This resulted in 24 trials within each block, which contained 18 common trials and 6 rare trials. Similar to Experiment 1, participants could complete a maximum of 5 blocks. After the first block, they were given a chance after completing each block to move straight to the test phase. The trial structure and response deadlines corresponded to Experiment 1.

In the test phase, participants were shown *incomplete* arrangement of geometric shapes and were asked to complete them. On each test trial, they were asked to select either **X** or **Y** to complete the arrangement. Similar to Experiment 1, each test item (incomplete arrangement of shapes) appeared 20 times. Similarly, the test phase was composed of 120 trials presented across 5 blocks of 24 trials.

Analysis and Exclusion We applied the same analysis and exclusion methods as in Experiment 1.

Results and Discussion

After exclusion, 30 participants made it into our analysis. The group-level mean probabilities are shown in Table 4. The results are a qualitative and ordinal match to Experiment 1. Participants showed a clear common preference for stimuli A, $M_A = 0.77$, 95% HDI $[0.68, 0.87]$, $\text{BF}_{10} = 10,316.41$.

Table 4: Group-level mean probabilities for each stimulus presented during the test phase in Experiment 2 after exclusion.

	$P(common)$	$P(rare)$
A	0.78	0.22
AB	0.95	0.05
AC	0.09	0.91
B	0.92	0.07
BC	**0.35**	**0.65**
C	0.08	0.92

Participants also showed a reliable rare preference on ambiguous BC trials, $M_{BC} = 0.64$, 95% HDI $[0.53, 0.75]$, $\text{BF}_{10} = 4.09$ This gives us a sufficient amount of evidence to conclude that we have observed the IBRE.

Here, we further demonstrated that the IBRE can arise without experimental-design components that explicitly promote an error-driven process.

Discussion

In this study, we tested a central assumption of the most prominent theories of the IBRE. This central assumption was prediction error.

In our first attempt, we implemented an observational learning version of the canonical IBRE procedure. This meant that features and category labels appeared on the screen at the same time. Participants learned about categories by reading complete sentences that describe what symptoms people exhibit and what diseases they have. The experiment included no feedback and required no responses from participants during training. From a theoretical perspective, there was no opportunity for making an explicit error, therefore the condition for prediction error. Regardless, we observed the inverse base-rate effect. One limitation of this approach was that there are assumed causal relationships between features (symptoms) and labels (diseases). These relationships can predispose participants to make feature-to-label predictions, which will result in prediction error and attentional relocation.

In our second attempt, we further removed the causal relationship between features and labels by changing the stimuli and their presentation. Here, participants saw nothing but an arrangement of geometric shapes, where previous category labels were treated as features. There were no causal links between features and labels. When participants needed to complete incomplete arrangements of these shapes, they still exhibited a rare bias on *BC* trials. We still observed the IBRE.

In both experiments, the IBRE occured without any explicit detail in the experimental procedure that would result in prediction error. Therefore, any theorised error-driven process must be able to operate without explicit feedback. Most prominent theories and their corresponding formal specification rely on relocating attention in response to prediction error. They are unable to process the current experiment, because they are not designed to encode information presented without feedback or fit cued recall tasks.

The two experiments suggest that the necessary conditions to observe the IBRE are fewer than previously established. In Experiment 2, the only remaining conditions are the two uniquely predictive features, an overlapping feature, sequential presentation and the base rate. (Johansen et al., 2007) gave evidence that even the overlapping features are not necessary. In their design, common category exemplars were made up of two features, while rare exemplars only had a single feature - disjoint cues. This can mean that common exemplars need to be composed of at least two features, while the rare exemplar must have only one feature. If the rare exemplar has one uniquely predictive feature and another feature that is not shared with the other category, we do not observe the effect. This is the shared-cue effect (Kruschke, 2001a; Wills et al., 2014). One can hypothesise that the shared cue and the disjoint cue are sufficient conditions, but at least one of them is necessary. In addition, one can hypothesise that in the absence of a shared feature, common exemplars must contain more features than rare exemplars.

Compared to Johansen et al. (2007), we do not consider the asymmetric representation and the base-rate neglect as necessary conditions. From our point of view, they are theoretical constructs. They are not part of the experimental design. We consider them to result from the environmental conditions specified in the experiment required for the IBRE as opposed to being necessary/sufficient conditions for the IBRE.

One hypothesised way asymmetric representation is manifested is the attentional tuning of cognitive representation of category exemplars. This is not necessarily absent in our experiment, but is not directly tested. Our experiments do not give direct evidence against the role of attention in developing the asymmetric representation. Nonetheless, it must not happen through an error-driven process as conceptualised in most prominent theories. To further investigate this, the cued-recall IBRE procedure could incorporate eye-tracking to measure dwell time and order of information encoding. Even if we observe more and longer fixations to C relative to B, attention can still be allocated asymmetrically to drive the distinguishability of categories. Explanations do not need to invoke an explicit error-driven process.

- auxiliary phenomenon (Wills CIRP addition) - Things that any theory of IBRE should explain - current theories fall short - eye-tracking and attention -

Conclusion

In Experiment 1, we conducted a successful conceptual replication of Johansen et al. (2007), which gave evidence for the IBRE being independent of supervised learning procedures. In addition, Experiment 2 further suggests that the IBRE generalises beyond simple predictive-learning (Kruschke, 1996; Don et al., 2021) and decision-making (Johansen et al., 2007) paradigms. Theories of IBRE are inadequate to account for these findings, because of their inability to extend beyond supervised learning.

Open Science

We made available the two experiments written in javascript, the analysis code, the raw data, and all other supplementary materials. Experiment 1 is shared via OSF, and GitHub. Experiment 2 is shared via OSF and Github. The main repository that includes this manuscript and links to the materials for the two experiments can be found on GitHub.

References

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1–12.

Don, H. J., & Livesey, E. J. (2017). Effects of outcome and trial frequency on the inverse base-rate effect. *Memory & cognition*, *45*(3), 493–507.

Don, H. J., & Livesey, E. J. (2021). Attention biases in the inverse base-rate effect persist into new learning. *Quarterly Journal of Experimental Psychology*, *74*(4), 669–681.

Don, H. J., Worthy, D. A., & Livesey, E. J. (2021). Hearing hooves, thinking zebras: A review of the inverse base-rate effect. *Psychonomic Bulletin & Review*, *28*(4), 1142–1163.

Inkster, A. B., Milton, F., Edmunds, C. E. R., Benattayallah, A., & Wills, A. J. (2022). Neural correlates of the inverse base rate effect. *Human Brain Mapping*, *43*(4), 1370-1380.

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

Johansen, M. K., Fouquet, N., & Shanks, D. R. (2007). Paradoxical effects of base rates and representation in category learning. *Memory & Cognition*, *35*(6), 1365 - 1379. (00012)

Kalish, M. L. (2001, June). An inverse base rate effect with continuously valued stimuli. *Memory & Cognition*, *29*(4), 587-597. (00018) doi: 10.3758/BF03200460

Kruschke, J. K. (1996). Base Rates in Category Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 3-26. (00218)

Kruschke, J. K. (2001a). The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1385.

Kruschke, J. K. (2001b). Toward a unified model of attention in associative learning. *Journal of mathematical psychology*, *45*(6), 812–863.

Medin, D. L., & Edelson, S. M. (1988, March). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*(1), 68-85.

Morey, R. D., & Rouder, J. N. (2022). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-4.4)

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

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.