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

Lenard Dome (lenarddome@gmail.com)
Brain Research and Imaging Centre
University of Plymouth, Research Way, Plymouth, PL6 8BU

Andy J. Wills (andy.wills@plymouth.ac.uk)
Brain Research and Imaging Centre
University of Plymouth, Research Way, Plymouth, PL6 8BU

Abstract

Include no author information in the initial submission, to facilitate blind review. The abstract should be one paragraph, indented 1/8 inch on both sides, in 9 point font with single spacing. The heading "**Abstract**" should be 10 point, bold, centered, with one line of space below it. This one-paragraph abstract section is required only for standard six page proceedings papers. Following the abstract should be a blank line, followed by the header "**Keywords:**" and a list of descriptive keywords separated by semicolons, all in 9 point font, as shown below.

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 posess 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 teir respective common and rare label. 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 Classical Probability Theory, the rational response is to attribute the common label to this ambiguous compound, because it is the most frequently occuring label. This rare bias on ambiguous combinations of BC have been observed across different variety of different experiments and manipulations (Kalish, 2001; Don & Livesey, 2017, 2017; Inkster, Mitchell, Schlegelmilch, & Wills, 2022; Wills, Lavric, Hemmings, & Surrey, 2014). For a more thourough introduction into this irrational bias, see a review by Don and Livesey (2021).

Assumptions of models 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 models are EXIT (Kruschke, 2001), 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. 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.

Current Study

In this work, we intend to test this basic assumption of models of the IBRE. 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. Any presumption of causal relationship can inadvertantly relacote 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. Johansen et al. (2007) tried to observe the inverse base-rate effect both in a predictive-learning condition and in an observational learning condition, where category labels were presented with features 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 that facilitates attentional relocation. This attentional tuning in turn pushes responding towards the rare label. Their design was optimised to investigate the hypothesised assymetric cognitive representation of the two categories - one of the assumptions. As a result, the only instance when they observed a rare bias was when common features presented in compound during training were paired with a rare feature presented by itself during training. This provided evidence for the assymetric cognitive represetation hypothesised to develop during training. 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. 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 dissipate between them. There is no need to relocate attention to reduce errors, therefore attention will not bias responding towards the rare label. But stronger $C \rightarrow rare$ feature-label connections will result in a rare bias on BC compunds during test. 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 Johansen et al. (2007), who tried to remove the shared cue that pushed attention towards C during training and resulted in the assymetric representation. Here, we strictly focus on prediction error in relation to 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), but not the main text. There are no details about the procedure of this experiment, so we cannot make direct comparisons. The only information available are 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 recieved course credit for their participantion. 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. The experiment code is avalaible at OSF and GitHub. Participants completed the experiment on their personal computer. 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 individual 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 were compiled from an online repository of popular baby names[1]. We selected the 50 most popular male and female names from 2021.

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 test phase. In the training phase, they encountered descriptions of people, the symptoms they experienced, and their respective disease. These descriptions appeared in the format of '*John has fever and rash, which belongs to disease Z*'. Partipants 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

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

a maximum of 5 blocks they could complete.

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

Exclusion To match performance with the predictive learning implementations of the IBRE, we decided to exclude participants whose test performance on the training items were below 0.75 accuracy. We arrived at this threshold by testing all difference levels of accuracy through 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.

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 rare 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 then 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).

Results and Discussion

After exclusion, 125 participants made it into our main analysis.

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

	$P(common)$	$P(rare)$
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

$M = 0.67$, 95% HDI $[0.62, 0.72]$, $\text{BF}_{10} = 1.11 \times 10^7$

Experiment 2

Method

Participants Recruited 65 participants.

Stimuli

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

Results and Discussion

30 made it into the analysis

Table 4: Caption.

	$P(common)$	$P(rare)$
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

$M = 0.64$, 95% HDI $[0.53, 0.75]$, $\text{BF}_{10} = 4.09$

Discussion

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

Open Science

Acknowledgments

In the **initial submission**, please **do not include acknowledgements**, to preserve anonymity. In the **final submission**, place acknowledgments (including funding information) in a section **at the end of the paper**.

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.

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. (2001). 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.