

Errorless irrationality: removing prediction errors from the inverse base-rate effect

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

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 unique combinations. People optimally label uniquely predictive features, B and C, presented individually with their 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 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 have been observed across different varieties of different experiments and manipulations (Kalish, 2001; Don & Livesey, 2017, 2017; Inkster, Mitchell, Schlegelmilch, & Wills, 2022; Wills, Lavric, Hemmings, & Surrey, 2014). For a more thorough 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. 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 inadvertently relocate attention in line with the direction of causality between features and labels.

Related Work

We will build on work carried out by Johansen et al. (2007). Johansen et al. (2007) failed to observe the inverse base-rate effect in an observational learning condition, where category labels were presented with features at the same time to participants. They largely interpreted their results in opposition to EXIT (Kruschke, 2001), but their experimental procedure does not correspond to the standard inverse base-rate effect procedure. Their design involves disjoint-cues (where categories shared no features in common), while the canonical design depends on a shared feature during training that facilitates attentional relocation, which in turn moves respond-

ing towards the irrational rare preference. 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. A simple explanation for that result is 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 push responding towards the rare label. Their design is therefore not sufficient to answer our question, whether prediction error is a necessary condition.

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.

Method

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

Results and Discussion

Experiment 2

Method

Table 2: 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

Results and Discussion

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

- 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., 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.
- 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.