

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

Lenard Dome (lenarddome@gmail.com)
Andy J. Wills (andy.wills@plymouth.ac.uk)

School of Psychology
University of Plymouth, Research Way, Plymouth, PL6 8BU

Abstract

The inverse base-rate effect is a robust irrational bias that arises when people face ambiguity. The most prominent theories of this irrational bias depend on prediction error. In this study, we gradually removed elements of a predictive learning design to test the extent to which error-driven processes underlie this bias. In our first experiment, we removed explicit feedback by implementing the inverse base-rate effect in an observational learning procedure. In our second study, we further removed any causal relationship between stimulus features and category labels by moving towards an unsupervised learning procedure. This removed any information participants could use to identify category labels. In both experiments, the inverse base-rate effect persisted and remained robust. This outcome suggests that this irrational bias is independent of supervised learning procedures. We propose that any theories and models of the inverse base-rate effect must manage information encoding and connection updates without explicit prediction error. We end by proposing two clear paths for future investigations.

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 category labels. These sets share a single feature, *A*, and possess a unique feature, *B* and *C*, predictive of their respective category label. The training thus can be summarised under two trial types, which we will express as $AB \rightarrow \text{common}$ and $AC \rightarrow \text{rare}$. During learning, these sets of features occur at different frequencies. The features under the common label usually occur three times as often as features under the rare label (Kruschke, 1996). Following training, people categorise features presented by themselves and in novel combinations. People tend to optimally label uniquely predictive features, *B* and *C*, with their respective common and rare labels when presented by themselves. Responses on the shared feature *A* tend to show the base-rate following, $A \rightarrow \text{common}$. But when uniquely predictive features are paired, *BC*, people tend to respond with the rare category label. According to Classical Probability Theory, the rational response is to categorize this ambiguous combination under the common label, because it is the most frequently occurring label. This rare bias on ambiguous combinations of *BC* has been observed across a 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 IBRE involve an attentional mechanism that drives both learning and responding. These theories are formal models. They are: a neural network with 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 relocates 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, but they are still learning to label the *AC* compound. The presence of *A* tends to push participants to generalize what they learned about *AB*, so they label *AC* as common, which results in an error. After making this 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, this attentional allocation persists and thus *C* will dominate responding. This results 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 creates an asymmetric cognitive representation that can be summarized as *AB* belongs to common, $AB \rightarrow \text{common}$, and *C* belongs to rare, $C \rightarrow \text{rare}$ (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, 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 might 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 standard trial-by-trial IBRE procedure without explicit feedback. In terms of a clear observational-learning version of the IBRE, Johansen, Fouquet, and Shanks (2007) included the result of a short pilot experiment in their Appendix. Unfortunately, there is no statistical analysis confirming that the IBRE is reliably observed. Johansen et al. (2007) report a sample size of 16. If we use an effect size of $d = 0.46$ observed by Inkster, Milton, Edmunds, Benattayallah, and Wills (2022) and an α of 0.05 with a non-directional alternative hypothesis, the experiment has 24% power¹. Given this information, this pilot experiment is underpowered. There are also no details about the procedure of this experiment. Therefore, we cannot make direct comparisons.

Nonetheless, Johansen et al. (2007) demonstrated that the inverse base-rate effect can occur without the traditional predictive learning design. In one of the conditions in their Experiment 3, the canonical inverse base-rate design (including the shared cue) was implemented in a list format. In this format, the trial-by-trial presentation of training items was turned into a list of 12 items fitted on a single page. Subsequently, participants made judgements about new cases on a separate page. In this condition, participants still exhibited a rare preference on *BC* trials. In another condition of Experiment 3, participants received the information about outcome frequencies as a summary before testing. This summary was presented as prose. After learning about feature-label information in this manner, participants did not show the IBRE but was matching the base rate. These experiments give us evidence about another boundary condition for the IBRE - itemized rather than summarized presentation of training items.

Additionally, there are at least three studies which directly look at error-driven processes in the IBRE. Don, Beesley, and Livesey (2019), demonstrated that on *AC* trials, people fixated on *C* longer than on *A* both pre-responding during stimulus presentation and post-responding during feedback (see also Kruschke, Kappenman, and Hetrick (2005)). This fixation bias increased with more training. They also observe greater fixation on *C* on *AC* trials, relative to *B* on *AB* trials. Furthermore, Wills et al. (2014) in an EEG study observed posterior selection negativity and concurrent frontal positivity for *C* relative to *B*, which gave evidence for an error-driven selective attentional learning process. These studies gave evidence that attentional reallocation occurs in line with the mechanisms of EXIT-like models. Inkster, Milton, et al. (2022) carried out a direct investigation into brain regions underlying error-driven learning in the IBRE. Their region of interest (ROI) analysis explicitly targeted areas that were hypothesized to be in-

involved in the computation of prediction error. They showed that these areas exhibited greater activation during the test phase for *C* relative to *B*. Given these findings, it is reasonable to suggest that prediction-error-driven attentional reallocation occurs in a standard supervised learning paradigm and it is driven by prediction error.

Experiment 1

Below, we detail our first attempt to test whether we could observe the rare response bias to *BC* 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, Experiment 1 will present category labels simultaneously 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. 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 names². 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 summarizes 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

¹We used the method provided by the R package pwr (Champely, 2020) to calculate power.

²The list was taken and later curated from a GitHub repository: <https://github.com/aruljohn/popular-baby-names>.

5 seconds. If the 5 seconds threshold was 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. A prompt appeared saying that *'Now you have the option to skip the rest of the training phase and move straight to the test phase. If you think you need some more time, you can continue training and study more patients.'* 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. Symptoms appeared in a sentence, such as *'John has a fever.'*, with a prompt asking participants to say what disease the person has, *'Does the patient have 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. So, the test phase 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
$AB \rightarrow \text{common}_1$ (x 3)	A, B, C,
$AC \rightarrow \text{rare}_1$ (x 1)	AB, AC, BC x 20

Analysis In order to test for 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, $P(\text{rare}|BC)$, 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 $P(\text{rare}|BC)$ is higher than 0.5, we conclude that we observed the IBRE. Values lower than 0.5 would indicate base-rate following. 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. This level of accuracy was the lowest at which the evidence that the participant performed better than chance was above the Bayes Factor of 3. We calculated the

Bayes Factor for binomial proportions via the method implemented in BayesFactor R package (Morey & Rouder, 2022).

Results and Discussion

After exclusion, 125 participants made it into our main analysis. In 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], $BF_{10} = 2.45 \times 10^7$. For this cue, people explicitly followed the base rate - responded rationally according to Probability Theory. In contrast, participants showed a reliable rare preference for BC, $M_{BC} = 0.67$, 95% HDI [0.62, 0.72], $BF_{10} = 1.11 \times 10^7$. 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.

	$P(\text{common})$	$P(\text{rare})$
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

Thus the current study strongly confirms that the IBRE can be observed in an observational procedure. In the current experimental design, the IBRE emerged in the absence of an explicit prediction error that drives the development of attentional allocation. All EXIT-like 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, EXIT-like theories cannot predict the presence of the IBRE.

One aspect of the current design is that participants might still experience internally-generated prediction errors from feature to categories 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 but a silent 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.

In Experiment 2, we address this by removing any design component that makes it clear to participants what the cate-

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



gory label is. And we also use stimuli 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 memorize the arrangement of these shapes. On each trial, we randomized the position of the geometric shapes in the arrangement. This further minimized the chances of having any design component suggestive of which feature is the category label.

Method

Participants We recruited 171 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 common and rare. 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 test phase. In the training/encoding phase, participants were repeatedly exposed to the exemplars and were asked to memorize 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* arrangements 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. Various arrangement of shapes appeared in the middle of the screen. The response options X and Y with the corresponding shapes were shown below. The prompt asked participants to pick one of the shapes to complete the arrangement. Participants could respond by pressing either X or Y on the keyboard. 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, 86 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, $M_A = 0.78$, 95% HDI [0.73, 0.83], $BF_{10} = 5.37 \times 10^{13}$.

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

	$P(\text{common})$	$P(\text{rare})$
A	0.78	0.22
AB	0.95	0.05
AC	0.08	0.92
B	0.92	0.08
BC	0.27	0.73
C	0.07	0.93

Participants also showed a reliable rare preference on ambiguous BC trials, $M_{BC} = 0.73$, 95% HDI [0.67, 0.79], $BF_{10} = 8.12 \times 10^8$. 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 that the IBRE is caused by the presence of prediction error.

In our first experiment, 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 described what symptoms people exhibited and what diseases they had. 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. Nevertheless, 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 might predispose participants to make feature-to-label predictions, which could result in prediction error and attentional reallocation.

In our second experiment, 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 were asked to complete incomplete arrangements of these shapes, they still exhibited a rare bias on BC trials. We still observed the IBRE.

The two experiments together 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. One hypothesized way asymmetric representation is manifested is the attentional tuning of cognitive representation of category exemplars. This is not necessarily absent in our experiments but is not directly tested. Our experiments do not give direct evidence against the role of attention in developing asymmetric representation or in its contributions to the emergence of the IBRE. Nonetheless, it must not happen through an error-driven process as conceptualized in the most prominent theories of the IBRE. To further investigate this, the cued-recall procedure could incorporate eye-tracking to measure dwell time and order of information encoding. EXIT-like theories can informally predict longer fixations on C relative to B during training, but it is unclear what mechanism underlies this attentional allocation without an explicit error-driven process. In addition, brain imaging could further elaborate on the overlap of activations between cued-recall and supervised learning procedures. This would enable us to pinpoint the networks that uniquely underlie this rare preference but are independent of task demands.

In both our experiments, the IBRE occurred without any explicit detail in the experimental procedure that would result in prediction error. Therefore, any theorized error-driven pro-

cess 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 accommodate the current experiments because they are not designed to encode information presented without feedback. Our results suggest that there could be a secondary cause of the IBRE not captured by previous process models.

Alternative Theories without Prediction Error

There are alternative theories of the IBRE that do not rely on processes that calculate prediction error. A version of the dissimilarity-similarity generalized context model DGCM, Stewart and Morin (2007) modified by O'Bryan, Worthy, Livesey, and Davis (2018) has been proposed as an explanation of the IBRE. From the perspective of DGCM, the main processes behind the rare preference are a combination of attention, memory strength of exemplars and dissimilarity from exemplars (stored category representations). $BC \rightarrow \text{rare}$ arises due to the combination of the following factors: the high salience of C relative to B and the impact of the dissimilarity of BC to the most remembered common category exemplar on the decision process. Nonetheless, the model can accommodate these results only informally because it does not specify the mechanisms which encode information and produce the attentional values of each stimulus. DGCM is a model of the test phase. In that sense, the current experimental manipulations cannot be represented in the model. This is the same challenge we encountered with process models - the specifications of these theories are unable to incorporate the changes to the experimental procedure.

Another alternative explanation is an eliminative-inference model (ELMO, Juslin, Wennerholm, & Winman, 2001). This approach considers the $BC \rightarrow \text{rare}$ bias to result from rule-learning and post-training inferential processes during BC trials. The process is most similar to *strategic guessing* (Kruschke & Bradley, 1995). Because of the dissimilarity of BC to the most frequently applied rule, $B \rightarrow \text{common}$, participants use the most similar rule applicable to BC from a "guessing set", $C \rightarrow \text{rare}$. This results in the $BC \rightarrow \text{rare}$ bias. Informally, ELMO could accommodate the observational learning paradigm, because of the presence of feature-to-label causality. Participants could extract the same rules hypothesised to underlie the rare bias in the standard procedure. Due to this presumed causality, ELMO could hypothesise that people encode rules about symptoms and diseases, which will similarly result in $BC \rightarrow \text{rare}$ during test. But it is unclear how it could deal with the cued-recall implementation, as there is no clear-cut feature-to-label relationship in the stimuli presentation that drives rule formation. In addition, ELMO also predicts the IBRE in the absence of a shared cue, even though humans do not show the IBRE under those conditions (Kruschke, 2001a).

Conclusion

Across two experiments, we investigated whether the demonstration of the IBRE requires the prediction-and-feedback components of the standard experimental procedure. 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 generalizes beyond simple predictive-learning (e.g. Medin & Edelson, 1988; Kruschke, 1996; Wills et al., 2014) and decision-making (Johansen et al., 2007) paradigms. This further suggests prediction error in terms of explicit feedback is not a necessary condition. Theories of IBRE are inadequate to account for these findings, largely because of their inability to extend beyond supervised learning.

Open Science

We have made available the two experiments written in javascript, the analysis code, the raw data, and all other supplementary materials. Experiment 1 is shared via <https://osf.io/auwvt/>, and <https://github.com/lenarddome/ply216-observational-ibre>. Experiment 2 is shared via <https://osf.io/2tmc4/> and <https://github.com/lenarddome/ply222-non-causal-ibre>. The main repository that includes this manuscript and links to the materials for the two experiments can be found on <https://github.com/lenarddome/pu093-errorless-attentionless-ibre>.

Acknowledgement

The data for Experiment 1 was collected as part of a final-year undergraduate project. We thank for the help and contributions of Selena Lockett and Charlotte Hubbard in the data collection for Experiment 1.

References

- Champely, S. (2020). pwr: Basic functions for power analysis [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=pwr> (R package version 1.3-0)
- 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., Beesley, T., & Livesey, E. J. (2019). Learned predictiveness models predict opposite attention biases in the inverse base-rate effect. *Journal of Experimental Psychology: Animal Learning and Cognition*, 45(2), 143.
- 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)
- 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(3), 849.
- 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.
- Kruschke, J. K., & Bradley, A. L. (1995). Extensions to the delta rule for associative learning. *Indiana University Cognitive Science Research Report*, 141. (00007)
- 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(5), 830.
- 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)
- 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.
- Stewart, N., & Morin, C. (2007). Dissimilarity is used as evi-

dence of category membership in multidimensional perceptual categorization: A test of the similarity–dissimilarity generalized context model. *The Quarterly Journal of Experimental Psychology*, 60(10), 1337–1346.

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.