2023

Testing the adequacy of formal models of an irrational learning effect

Dome, Lenard

For my Dad.

ʾinna lillahi wa ʾnna ʾlayhi rağiʿun

إِنَّ لِلَّهِ وَ انَّ اليهِ رَجِعُن

UNIVERSITY OF
PLYMOUTH

TESTING THE ADEQUACY OF FORMAL MODELS OF AN IRRATIONAL LEARNING EFFECT

by

LENARD DOME

A thesis submitted to the

University of Plymouth

in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology

September 20, 2023

Acknowledgements

In an expression of true gratitude, sadness is conspicuous only by its absence.

– *Marcus Aurelius*

I have been incredibly fortunate to pursue a PhD. I am even more fortunate to be surrounded by such wonderful and supportive people.

First, I am grateful to my supervisor, Andy Wills, for lending his expertise, wisdom, time, and endless patience. He is probably the only person who doesn't flinch when I bring up ideas from science-fiction (specifically Psychohistory) during a formal meeting. I am eternally thankful to you.

Second, I am grateful to my mother, Maria, my sister, Lucia, and my late father, Janos. Without them, I would never have been able to pursue science as a vocation. They encouraged me and stood by me no matter what.

I am also grateful to my wife and partner-in-crime, Kenza. Millions of words cannot convey how grateful I am for you and for your support. Thank you for laughing with me through tough times and always being the sharpest knife in the drawer.

A lifetime is not enough to repay the debt I owe to these people.

I would also like to thank Rory Baxter for sharing my weird film choices, which caused us to watch three Carpenter movies on the same day. Also, thank you for introducing me to Elden Ring. I have never recovered from the Lands Between. I would also like to thank Karol Nedza for our deep-dives into Philosophy and Russian Literature. We had some of the best conversations over coffee on Analytic Philosophy, Dostoyevsky, and Stoicism. Nothing could replace it, Comrade. I would also like to thank Stuart Spicer, Marius Golubickis, Elliot Walby, Tara Zaksaite, Paul Sharpe, Rory Spanton, and Leonie Cooper.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee. Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

Relevant scientific seminars and conferences were regularly attended at which work was often presented:

Experiment 1 was an already existing unpublished data set in the lab at the time of starting the research degree. The data is presented here to maintain narrative coherence.

Word count for the main body of this thesis: **26625**

Signed: 

Date: 9/20/2023

Publications:

Dome, L. & Wills, A.J. (*preprint*) Better Generalization Through Distractions. *PsyArxiv*. DOI: https://doi.org/10.31234/osf.io/eskr9.

Dome, L. & Wills, A.J. (*under review*) g-distance: On the comparison of model and human heterogeneity. DOI: https://doi.org/10.31234/osf.io/ygmcj.

Dome, L. & Wills, A. J. (2023) Errorless irrationality: removing error-driven components from the inverse base-rate effect paradigm. *Proceedings of the 45th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.
URL: https://escholarship.org/uc/item/0kw671vv.

Presentations and conferences attended:

2023 45th Annual Conference of the Cognitive Science Society. Sydney, Australia.

2023 MathPsych/ICCM/EMPG: Amsterdam, Netherlands.

2023 Experimental Psychology Society Meeting: Plymouth, UK.

2022 22nd Associative Learning Symposium: Gregynog Hall, Wales, UK.

Software:

psp maintainer, developer, author. GitHub.

catlearn senior developer. GitHub.

clobe author, maintainer, developer. GitHub.

Testing the Adequacy of Formal Models of an Irrational Learning Effect

Lenard Dome

Abstract

The inverse base-rate effect (IBRE) is an irrational phenomenon in predictive learning characterized by as a preference for rare, unlikely outcomes in the face of ambiguity. This thesis investigates the adequacy of formal explanations for this puzzling phenomenon. In the first project, I will focus on mechanisms of learning that mathematical models posit underlie this preference. A class of attentional explanation produces a counter-intuitive prediction: the effect disappears under concurrent load. I confirm the prediction, but only when participants were under an obvious time constraint – irrationality reduces under increased task demands. This suggests that multiple learning mechanisms operate independently and are differentially affected by concurrent load. In the second project, I test basic assumptions of the most prominent theories: this irrational bias depends on prediction error. Here, I gradually removed elements of a predictive learning design to test the extent to which error-driven processes underlie this bias. Throughout my attempts, the inverse base-rate effect persisted and remained robust. This outcome suggests that this irrational bias is independent of supervised learning procedures - a big change in the problem structures of the IBRE. In the third project, I look for the most adequate formal computational model of the canonical IBRE. In addition to group-level accommodation, I also incorporate heterogeneity into the benchmark. To accomplish this, I developed g-distance, which incorporates the extent to which models exhibit a similar range of behaviors to the humans they model. Applying it to five models of the IBRE reveals that none of the models outperform a random model. While analyzing the human data, I also discovered that the group-level result was observed in less than 1% of individuals. These projects provide new insight into the IBRE and how we should approach building and evaluating models of the IBRE and associated phenomena. I will discuss these insights in detail and how they influence future research on the IBRE.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When you hear hoofbeats, think horses, not zebras. This adage is sometimes used in medical education to make the following point: if the symptoms are ambiguous, a common disease is a more appropriate diagnosis than a rare one. This seems uncontroversial – in the absence of information to the contrary, the most common outcome is the most likely. People tend to endorse this principle but sometimes act in opposition to it. The register of human errors is full of systematic biases and mistakes in judgments (Tversky & Kahneman, 1974; Tversky, Slovic, & Kahneman, 1990), pseudo-contingencies without correlation data (Chapman, 1967; Hamilton & Gifford, 1976; Fiedler, Kutzner, & Vogel, 2013), and the tendency to neglect prior experience under uncertainty (Bar-Hillel, 1980; Krynski & Tenenbaum, 2007).

1.1 The Inverse Base-Rate Effect (IBRE)

One particularly striking example of this tendency is observed in the *inverse base-rate effect* (IBRE) (Medin & Edelson, 1988). In a typical IBRE procedure, people learn through a series of examples that a combination of two symptoms leads to one disease ($AB \rightarrow 1$), while a different but overlapping combination of symptoms leads to a different disease ($AC \rightarrow 2$). Crucially, the two diseases occur at different frequencies, with disease 1 typically occurring three times as often as disease 2. When subsequently asked about the ambiguous case of symptom A on its own, people tend to predict disease 1, the most frequently-seen outcome. This response seems rational; A was followed by both disease 1 and disease 2, but the common outcome, disease 1, occurred three times as often, so disease 1 seems more likely overall. In terms of Bayes' Theorem, we can express this more formally:

$$P(1|A) = \frac{P(A|1) \times P(1)}{P(A)}, \tag{1.1}$$

where $P(1|A)$ is the probability of disease 1 given symptom A; and

$$P(2|A) = \frac{P(A|2) \times P(2)}{P(A)}, \tag{1.2}$$

where $P(A|2)$ is the probability of A given disease 2. Using these formulas, we can express the likelihood of diseases 1 and 2 given symptom A as follows:

$$\frac{P(1|A)}{P(2|A)} = \frac{P(A|1) \times P(1)}{P(A|2) \times P(2)}. \tag{1.3}$$

Here, note that the $P(A|1) = P(A|2) = 1$ because the presence of symptom A preceded all instances of disease 1 and 2. There were no trials without A during training. $P(1)$ occured three times as often as $P(2)$ occured throughout training, so that $P(1) \div P(2) = 0.75/0.25 = 3$. Bayes' Theorem posits that disease 1 is three times as likely as disease 2 given symptom A. So participants match base-rate and predictions of probability theory.

However, when presented with BC, people tend to predict disease 2 - the rarer disease. Like A, BC is also ambiguous; symptoms B and C have uniquely and perfectly predicted their respective outcomes. However, in this case, people predict the rare disease rather than the common one. This generalization opposes the base rates of the diseases and thus arguably irrational. According to Classical Probability Theory, the rational response is to categorise this ambiguous combination under the common disease 1, because it is the most frequently occurring outcome.

$$\frac{P(1|BC)}{P(2|BC)} = \frac{P(BC|1) \times P(1)}{P(BC|2) \times P(2)} \tag{1.4}$$

In our case, $P(B|1) + P(C|1) = 1 + 0$ so that $P(B \cap C|1) = 1$[1]; and $P(B|2) + P(C|2) = 0 + 1$ so that $P(B \cap C|2) = 1$. This means that $P(BC|1) = P(BC|2) = 1$. $P(1)$ and $P(2)$ are the same as in the case of symptom A, so the right-hand side of the equation evaluates to 3. Therefore classical probability predicts that disease 1 is three times more likely than disease 2 given BC. Participants nonetheless exhibit a rare bias, which is in direct contrast to the overall base-rate and the prediction of classical probability theory. This rare bias on ambiguous combinations of BC has been observed

[1] In this notation, $\cap$ denotes the intersection of events, which is when B and C occurs together.

independently across a variety of experimental manipulations (Kalish, 2001; Don & Livesey, 2017, 2021; Inkster, Mitchell, Schlegelmilch, & Wills, 2022a; Wills, Lavric, Hemmings, & Surrey, 2014; Shanks, 1992; Johansen, Fouquet, & Shanks, 2007, 2010; Kalish, 2001; Sherman et al., 2009). For a more thorough introduction to this irrational bias, see an excellent review by Don, Worthy, and Livesey (2021).

Initially, Medin and Edelson (1988) framed their study as a continuation of the argument drawn by McClelland and Rumelhart (1985) that stresses an important point: understanding decision-making requires the study of the mechanisms of learning in addition to the structure of the problem domain. The IBRE initially served to demonstrate this. First, base-rate information in the IBRE is conveyed through experience, extracted by the learning processes, and represented in memory. This results in a non-uniform generalization of base-rate information - participants match the base-rate on some cues and ignore the base-rate on others. Second, at the time, demonstrations of how people use, misuse, and combine prior odds (base-rates) mostly came from the type of pen-and-pencil tasks popularised by Tversky and Kahneman (1974, 1980, 1981). Experience, learning, and generalisation played little role in this strain of base-rate research. As a matter of fact, Johansen et al. (2007) later showed that this irrational preference for rare on *BC* disappears when the problem takes the form of a traditional pen-and-pencil test. In that implementation, the task directly communicates base-rate information to participants in ways similar to the classical cab problem (Tversky & Kahneman, 1980). Medin and Edelson (1988) and Gluck and Bower (1988) are examples of the minority of investigations from that time. They both looked at how base-rate knowledge is acquired and later non-uniformly applied to novel events. This minority group of investigations stressed the importance of examining the combination of learning mechanisms and the problem structure. As it happened, models of learning designed for this exact problem structure became the most successful theories of the IBRE.

1.2 Theories of IBRE

The most successful explanations of the IBRE propose that it is caused by error-driven reallocation of attention (Kruschke, 2001b, 2003; Paskewitz & Jones, 2020). Participants first learn $AB \rightarrow 1$, because this occurs most frequently. When encountering the rarer symptom combination, *AC*, they generalize what they have learned about the common cue, $A \rightarrow 1$, to *AC*, leading to an error. In order to avoid this error in the future, they shift their attention away from *A* and towards *C*. This increased attention to *C* persists over time. Thus, when the participant is presented with the

novel and ambiguous combination *BC*, they attend to *C* more than *B*, leading to *C* dominating responding, hence a preference for outcome 2 - the rarer disease.

There is substantial behavioural and neuroscientific evidence for this attentional re-allocation account. Don, Beesley, and Livesey (2019) used eye-tracking to demonstrate that, during training, participants fixated longer on *C* than *A* on *AC* trials, relative to *B* on *AB* trials (see also Kruschke, Kappenman, and Hetrick (2005a). Wills et al. (2014), using an EEG-ERP methodology, demonstrated *C* was more attended than *B* at test. Inkster, Milton, Edmunds, Benattayallah, and Wills (2022), using fMRI, provided further evidence for an error-driven attention account– brain areas associated with prediction error showed greater activity for *C* relative to *B*.

This idea of error-driven attention has been formally captured in a multitude of mathematical models (Mackintosh, 1975; Kruschke, 1996; Paskewitz & Jones, 2020). The most famous in this family of models applied to these results is the model of EXemplar-mediated attention to distinctive InpuT (EXIT Kruschke, 2001b). EXIT originates from (Kruschke, 1992) and combines an error-driven learning assumption of Gluck and Bower (1988) and the exemplar modelling of category learning of Nosofsky (1986). Both Gluck and Bower (1988) and (Nosofsky, 1986) have their own historical roots. Gluck and Bower (1988) is an extension of Rescorla and Wagner (1972)'s delta-rule (or least mean squares) model of associative learning, which is identical to the learning algorithm of Widrow and Hoff (1960).[2] Nosofsky (1986) came after two independent publications of a formal exemplar theory of categorization in the same year by Medin and Schaffer (1978) and (Brooks, 1978). EXIT (Kruschke, 2001b) employs attentional shifts through differentially weighted cue-category connections during both learning and responding, which gives rise to irrational cue utilization after the learning has been completed. Here, we briefly describe the internal operations of EXIT, but for a more thorough description, see Appendix B. For the mathematical specification, see Kruschke (2001b).

In EXIT, decisions are based on the sum of exemplar activations under each category. Each exemplar activates according to its similarity to the current input, where dimension-specific attentional tunings of each exemplar weigh in heavily on these similarities. The choice rule simply converts these category activations into choice probabilities. Learning in the model is based on prediction error as calculated by a summed error term. Attention is similarly adjusted via gradient descent on error and reiterates within each trial. The overall optimisation problem the model tries to solve is the reduction of errors - maximising accuracy during learning. As a result, distinctive features (the

[2]Interestingly, Widrow and Hoff (1960) and Rescorla and Wagner (1972) published their algorithms more than a decade apart, independent of each other.

ones that are most useful in solving this optimisation problem) acquire higher attentional values because they are uniquely predictive of their category. But an increase in attention to one feature will lead to a decrease in attention to other cues. This decrease lowers attentional salience for unpredictive cues and increases attentional salience for predictive cues.

EXIT has been shown to outperform other models trying to explain the IBRE (Kruschke, 2001a, 2003; Don et al., 2019; Don & Livesey, 2021) and had a wide range of successes across many phenomena (Kruschke, 2011). It is often considered the most adequate of the available formal accounts. Nonetheless, Paskewitz and Jones (2020) argued that EXIT is a quite complex realization of the attentional-associative approach, which can be dissected into multiple, simpler, models of error-driven attentional reallocation. They presented a dissection of EXIT into its constituent processes, gradually building up towards the full formal specification of Kruschke (2001b). Paskewitz and Jones (2020) showed that the complexity of EXIT is not necessary to accommodate the base result of the IBRE $A \rightarrow common$ and $BC \rightarrow rare$. In that process, they discovered that rapid attentional shifts, exemplar-mediated attention and exemplar-mediated similarity are not necessary to reproduce the canonical base result, only error-driven global (not exemplar-specific) attention. The simplest model they presented that could still account for the IBRE was a four-layer neural network with competitive attentional gating and an exponential choice rule (softmax). For a more thorough description of the architectures of these models, see Appendix B. For the mathematical specification, see Paskewitz and Jones (2020). In what follows, I will discuss three topics that could challenge these formal explanations.

1.3 Mechanisms of learning

A critical aspect of these models is that they describe both how learning takes place and how post-learning processes apply the outcome of learning. In the models, learning abstracts away the basic properties of a set of events. Decision-making utilizes these abstractions (exemplars, rules, and attentional tunings). These approaches posit that decision- and categorization-related behaviour emerges from how psychological processes represent, manipulate and transform this information. In the case of the IBRE, learning extracts some properties of the set of experiences as exemplars and attentional tunings. This is done with a single-minded attempt to reduce errors. For example, C acquires higher attentional salience than A and B, because of the errors associated with AC during the learning of cue-outcome mapping. Then the decision mechanism operates on these partial activations of similarity to exemplars (AC, AB) and weighted attentional connection to out-

comes (high $C \rightarrow$ *rare*, low $A \rightarrow$ *common* and $B \rightarrow$ *common*). The model superimposes attentional tuning on cognitive representations, which results in the asymmetric cue-outcome representation hypothesized to underlie the IBRE (Kruschke, 2001a). These representations are $C \rightarrow$ *rare* and $AB \rightarrow$ *common*. They are composites of both weighted connections of cues and outcomes (often underpinned by similarity) and attentional connections of cues and outcomes. People repeatedly argued that what is essential to produce the IBRE is the attentional component of these systems (Shanks, 1992; Kruschke, 2003; Paskewitz & Jones, 2020). Therefore, one might argue that disruption of this attentional learning process during learning will also disrupt the IBRE.

1.4 Problem Structure

Another important aspect of all these accounts is that they are the exact right fit for the problem structure. Here, problem structure refers to choices in experimental design such as category structures, stimulus composition, the method of conveying information, and the presence or absence of feedback. Problem structure strongly relates to the problem of how an experiment or any computable problem is represented to the model (e.g. compound stimuli as integer vectors, continuous stimuli dimensions as double vectors, images as pixel matrices or multidimensional arrays). It also determines how generalisable a given model is across paradigms and phenomena. Understandably, mechanisms of learning are not independent of the model's target domain of problem structures.

Models are built to map to a particular combination of experimental design components - they are *not* task-agnostic. For example, ALCOVE (Nosofsky, 1986), a model of category learning, is built to accommodate problems where people sequentially label various stimuli and learn through immediate feedback. ALCOVE is not suited to solve free-classification problems such as (Medin, Wattenmaker, & Hampson, 1987), because its architecture is built to deal with different problem representations. Or consider SUSTAIN (Love, Medin, & Gureckis, 2004), a model that is fit to deal with both supervised learning (learning through immediate feedback after choices) and unsupervised learning (learning to impose structure on its experience in the absence of feedback). In this case, almost all demonstrations of the IBRE involves a traditional predictive learning design - see Johansen et al. (2007) for a notable exception that I will further discuss in Chapter 3. That is to say, experience accumulates bit by bit through sequential encounters of distinct events. People repeatedly make decisions and receive immediate feedback about the accuracy of those decisions. If we take the formal models literally, this guess-and-feedback component in the predictive learning procedures is necessary for the presence of the IBRE. The model needs explicit feedback in

order to compute prediction error. This prediction error is then used to determine predictive cues that will help the model to avoid errors. Attention will then be adjusted according to gradient descent on error, so predictive cues will be highly attended. Without the presence of a prediction error - as it is conceptualised within these models' architectures - IBRE will not arise. Therefore, changing the problem structure: tests how generalisable the IBRE is across various task formats; establishes what the necessary conditions for this irrational preference to arise are; determines how far experimental procedures can move from the target domain of these models without disrupting the irrational preference.

1.5 Heterogeneity in the inverse base-rate effect

While individual differences have a long history in the literature (Merrell, 1931; Sidman, 1952; Blyth, 1972; Estes, 1956; Siegler, 1987; Ashby, Maddox, & Lee, 1994; Heck et al., 2022), they had remained unstudied in the IBRE until this thesis. We did not know the extent to which individuals in the standard procedure exhibit different combinations of response tendencies as a function of the stimuli. However, there is substantial evidence that group-level averages hide significant variances in behaviour in other phenomena (He, Liu, Eschapasse, Beveridge, & Brown, 2022; Conaway & Kurtz, 2017; Nosofsky & Hu, 2022; Lee, Hayes, & Lovibond, 2018). Across multiple experiments, it has been observed that distinctly different results can accompany the rare preference on *BC* trials. For example, in the standard procedure, the most common result relating to predictive cues, B and C, is that B is less associated with the *common* outcome than C is with the *rare* outcome (Medin & Edelson, 1988; Kruschke, 1996). This is generally referred to as $B < C$. Nonetheless, in some cases, $B > C$, the exact opposite pattern, is also observed in combination with $BC \rightarrow$ *rare* (Bohil, Markman, & Maddox, 2005; Wills et al., 2014; Winman, Wennerholm, Juslin, & Shanks, 2005; Inkster, Mitchell, et al., 2022a) These are distinct group-level results that models of the IBRE must also accommodate. It is not controversial to hypothesize that this variance can expand into individual-level behaviour. Let us consider the following. If on the individual-level $B > C$ and $BC \rightarrow$ *rare* only occur with $A \rightarrow$ *rare* but not with $A \rightarrow$ *common*, then this unexpected combination of results would challenge what many of the models I discussed above predict. A has been associated with both outcomes, but it has occurred with the *common* three times as often than with *rare*. This is something models would arguably struggle to fit. It is important to investigate this variance because it gives us a unique opportunity to incorporate theoretically crucial empirical patterns masked by group-level averaging.

1.6 Structure of the Thesis

In this thesis, I present three threads of research connected by the IBRE that correspond to mechanisms of learning, problem structure, and heterogeneity. Chapter 2 will discuss and test a counterintuitive prediction of EXIT: IBRE disappears when people are distracted. Across two experiments, this prediction is tested by implementing a concurrent load procedure from Wills, Graham, Koh, McLaren, and Rolland (2011) and Seabrooke, Wills, Hogarth, and Mitchell (2019) in the inverse base-rate effect paradigm. Concurrent load allows investigation of how attentional and cue-outcome learning mechanisms operate under high task demands.

Chapter 3 will explore how much we can change the problem structure: explore how far we can stray from the traditional predictive learning design. All models assume that error-driven processes give rise to the IBRE. This will be investigated by gradually removing design components that could promote the explicit generation of prediction error. First, I will implement the IBRE paradigm as an observational learning task and later as a cued-recall memory task. This span of different experimental procedures allows the investigation of how dependent the IBRE is on the presence of an explicit prediction error. The overarching goal here is to find problem structures outside the model's scope but still suitable to give rise to the IBRE.

Chapter 4 will present a different type of approach and will primarily focus on computational modeling of the IBRE. In Chapter 4, I will develop a framework to assess how well models of the IBRE accommodate heterogeneity in a traditional and minimal IBRE procedure. I attempt to determine the most adequate model of the IBRE that also incorporates heterogeneity. This chapter involves the development of a model adequacy framework that generalizes beyond models of the IBRE. I discuss potential uses and extensions of this measure but also present new findings about heterogeneity in the IBRE and how well existing models accommodate it.

Chapter 2

Concurrent Load and the Inverse Base-Rate Effect

2.1 Introduction

In this Chapter, I will investigate how learning processes operate with a special emphasis on attentional learning - how attentional reallocates to reduce errors. The attentional reallocation account arguably provides the most complete account of the IBRE and associated phenomena (e.g. Kruschke, 2001a, 2003, 2011; Inkster, Mitchell, Schlegelmilch, & Wills, 2022b; Don et al., 2019; Don & Livesey, 2021; Paskewitz & Jones, 2020). EXIT also makes a clear, striking, and counter-intuitive prediction: distraction should reduce the size of the IBRE (See Figure 2.1, and Appendix A for the simulation details). More precisely, interference from concurrent load is assumed to reduce the rate at which attention is re-allocated in models like EXIT (Nosofsky & Kruschke, 2002). This impairment of attentional tuning leads to a reduction in the size of the IBRE. Thus, given that the IBRE is arguably a case of irrational generalization, EXIT predicts that *people will respond less irrationally when they learn about this task while distracted, relative to when they give it their full attention*.

At first glance, this striking prediction appears to have already been disconfirmed: Lamberts and Kent (2007) showed that the IBRE was unaffected by concurrent load. Although their analysis was based on null-hypothesis significance testing, a Bayesian re-analysis of their data confirms evidence for the null ($BF = 0.28$, see Appendix A). However, Lamberts and Kent (2007) 's participants learned under conditions of full attention, with concurrent load applied only after learning, during a test phase where participants made decisions without further feedback. In contrast,

EXIT's prediction about concurrent load concerns its application during learning, and eye tracking also shows that attentional re-allocation happens during learning (Don et al., 2019). It is also the case that, in other predictive learning tasks, applying concurrent load during learning affects behaviour more strongly than applying it during test (see Wills et al. (2011)). Thus, in the current studies, I examined the effects of concurrent load during training on the IBRE.

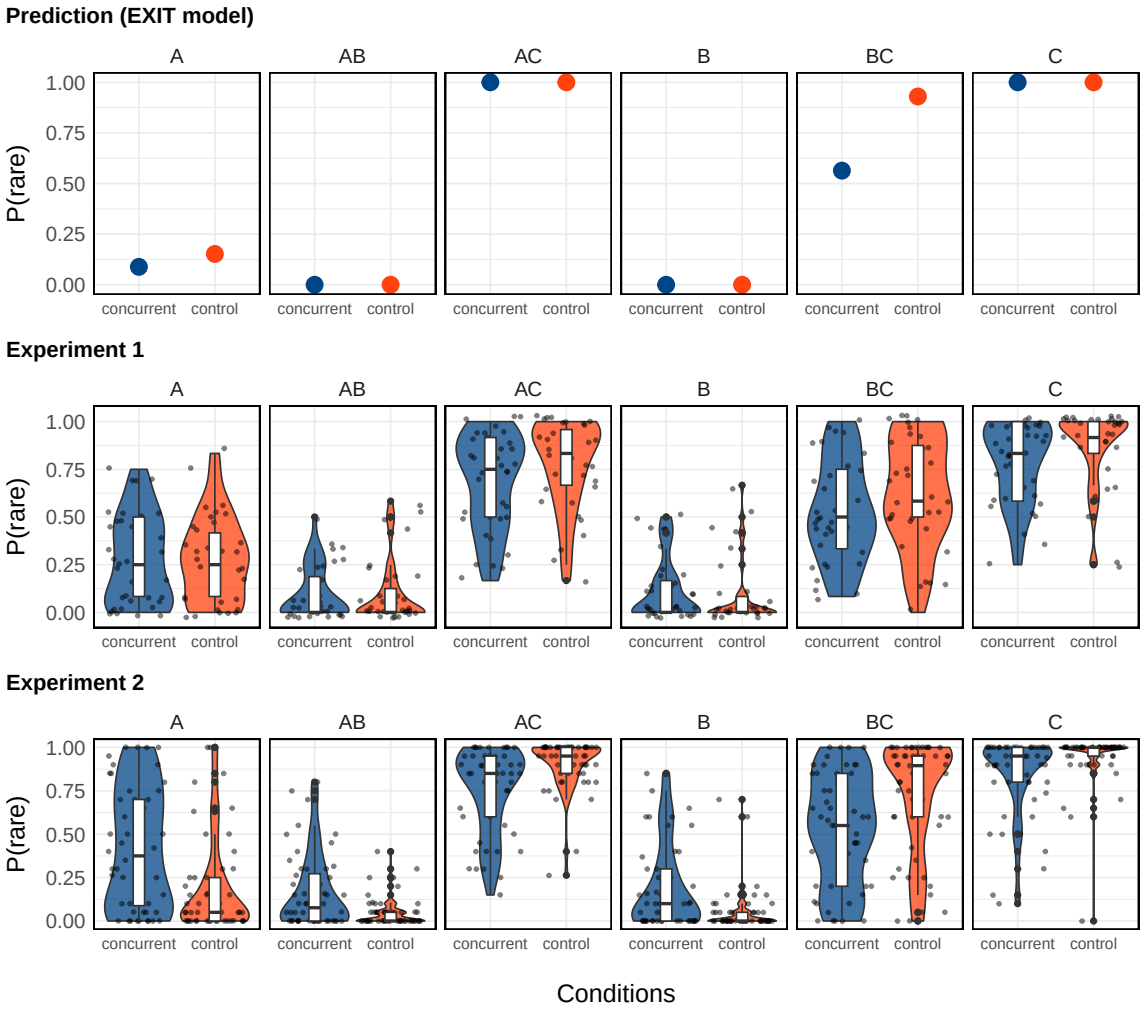


Figure 2.1: Response probabilities for all test items in the control (orange) and concurrent load (blue) conditions. For EXIT, the dots show the mean predicted response probabilities. For Experiments 1 and 2, each dot is a single participant. Distributional information is shown as a boxplot, a violin plot, and individual data points. The box plot shows median performance and interquartile range. The violin plot is a density plot, rotated through ninety degrees, and mirror copied to produce the symmetrical pattern showm, see Hintze and D. (1998) on violin plots. The response probabilities are shown for all test items that were present for all experiments and simulations.

Notes

Note that Experiment 1 was an already-existing data set in the lab and was conducted as a final-year research project before the commencement of the Ph.D. All analyses, findings, and conclusions reported here are novel and conducted as part of the Ph.D.

2.2 Experiment 1

Our first experiment differed from Lamberts and Kent (2007) in two principal respects. First, Lamberts and Kent (2007) applied load during test, while I applied it throughout the experiment. Second, they used a within-subjects design, while I opted for a between-subjects approach in order to eliminate the possibility of transfer effects.

2.2.1 Method

Participants and Sample Size

72 participants ($M_{age} = 20.12$[1]) took part in my experiment (36 in the concurrent load and 36 in the control condition). Participants were undergraduate students at Plymouth University, who completed the experiment for course credit. The experiment ran in person. The sample size was determined in advance of data collection and was sufficient to detect a medium-to-large effect (d = .67) at 80% power.

Apparatus and Materials

The experiment was implemented in Psychopy v2.7 (Peirce et al., 2019), with participants responding using a standard USB keyboard. Auditory stimuli were delivered through Behringer HPM1000 headphones.

The stimuli were the words "ear aches, skin rash, back pain, dizziness, sore muscles, stuffy nose", which were randomly assigned to abstract cues "A, B, C, D, E, F" for each participant. Traditionally, IBRE experiments use a "doubled-up" design that incorporates two sets of symptom-disease mappings. The two sets have the same structure where each has an overlapping symptom (A and D), a common symptom (B and E), and a rare symptom (C and F). The abstract stimulus types are presented in Table 2.1. The position of individual symptoms on the screen was counterbalanced. For example, if the stimulus consisted of "ear ache" on the top and "skin rash" on the bottom, it was also presented as "skin rash" at the top and "ear ache" at the bottom during the same session. The disease names were "Q, R, U, P", which corresponded with the buttons on the keyboard. The diseases were randomly assigned to common and rare diseases.

[1]The mean age is an estimate of the demographic information available from 42 participants at the time of writing this manuscript.

Table 2.1: Abstract trial types during the training and test phase of Experiment 1.

Training (Relative Frequencies)	**Test** (x6)
$AB \rightarrow common_1$ (x 3)	A, B, C, AB, AC,
$AC \rightarrow rare_1$ (x 1)	BC, ABC
$DE \rightarrow common_2$ (x 3)	D, E, F, DE, DF,
$DF \rightarrow rare_2$ (x 1)	EF, DEF

Concurrent load

I used a concurrent load procedure that was also employed in previous learning tasks (Wills et al., 2011; Seabrooke et al., 2019). At the beginning of the trial, participants were presented with a random list of six single digits through headphones before the categorization task. The numbers were presented at 330 ms intervals and were randomized at each presentation. After feedback on the categorization task, participants received a probe (a random number from the first five members of the six digits), and were asked to enter the number that followed the probe in the list from the beginning of the trial. For example, if the participant heard "2, 5, 4, 3, 9, 7", and the number probe "3" was presented, the correct answer was "9". No feedback was given and there was no time limit to respond.

Procedure

In the load condition, participants started with 10 practice trials for the concurrent load task. After this, they moved on to the training phase. Each trial began with a fixation cross displayed for one second. This was followed by six spoken digits, which lasted 2.37s. When the presentation of the six digits was concluded, participants were immediately presented with two symptoms, and they were asked to make a diagnostic choice ("Which disease do these symptoms belong to?"). Stimulus presentation was terminated when participants made a response. When they responded, feedback was given with either "CORRECT" or "INCORRECT" displayed on the screen. The correct disease name was also displayed. This feedback lasted for two seconds, after which a digit probe was presented. Participants were then asked to input the digit that followed the probe from the list of spoken digits presented at the start of the trial. After making the response, there was a 500ms inter-trial interval.

This training phase consisted of up to 10 blocks of trials. Each block contained 16 trials, with each of the two common diseases presented six times, and each rare disease twice, in random order. Participants were assessed against a learning criterion of 100% correct in one block, upon which

they were moved to the test phase. If they did not reach that criterion, the training continued until they completed the 10th block.

The test phase consisted of 84 consecutive trials, where participants were asked to diagnose new combinations of symptoms, along with the training items (see Table 2.1 for the abstract stimuli). The trial structure was the same as the training phase, except that participants didn't receive feedback after making a diagnosis. Instead, the message "No diagnostic feedback report available. Your response has been recorded" was displayed for two seconds. After this, participants completed the digit task, which was followed by a 0.5s inter-trial interval.

In the control condition, the trial structure was the same as the concurrent load condition, except in the following respects. Instead of spoken digits, participants were presented with a blank screen for 2.37s. In addition, participants completed a filler task instead of the concurrent load task that followed the diagnosis. In the filler task, participants were presented with a digit and were asked to press the corresponding numeric key on the keyboard. These changes were made to match the timing and response requirements of the load condition, while minimally loading the participant.

Analysis

Abstractly identical cues have been combined. For example, A represents (in the analysis below) responses to both A and D, BC represents responses to both BC and EF, etc. In order to test for the presence of the IBRE, I calculated the Bayes Factor for a paired comparison between rare-disease and common-disease responses on BC. This was tested against a null model of the differences, $\mu = 0$. To test the effects of concurrent load, I carried out a mixed-effects Bayesian ANOVA testing for the main effects of response options (common vs. rare) and concurrent load (control vs. concurrent), and for the interaction of response option and concurrent load. Following (Jeffreys, 1998), Bayes Factors above three were taken as evidence for a difference, while Bayes Factors less than one-third were taken as evidence for the null. All Bayesian analyses were conducted using the BayesFactor package (Morey & Rouder, 2018) of R (R Core Team, 2023).

2.2.2 Results and Discussion

I excluded one participant who made more than one mistake on the (10^{th}) block of training. This resulted in 35 people in the control and 36 people in the concurrent load condition. Figure 2.1 shows the probability of responding with rare for each test item in both conditions. In the control condition, participants showed a clear ($BC \rightarrow rare$) preference, ($\mathrm{BF}_{10} = 5.60$), replicating the

IBRE. In the concurrent-load condition, the evidence for a (*BC* → *rare*) preference was inconclusive, ($BF_{10} = 0.93$). While there was evidence for the null in the main effect of concurrent load, ($BF_{10} = 0.18$), the evidence for an interaction was inconclusive, ($BF_{10} = 0.54$). Thus, the results of my first experiment, while suggestive of an effect of concurrent load, were overall inconclusive.

A possible reason for this lack of certainty was the largely self-paced nature of the task. It has been previously found that where the imposition of a concurrent load task is accompanied by a reduction in time pressure, the latter can nullify the former (Newell, Moore, Wills, & Milton, 2013). Perhaps the absence of any substantial time pressure in Experiment 1 had a similar effect, with participants making use of the self-paced nature of the task to compensate, at least partially, for the additional demands of the concurrent load task. If so, they might still have had time to partially reallocate attention during training. I address this possibility in Experiment 2 by imposing a shorter time limit in both the control and the load conditions. I predict this will reduce people's ability to compensate for the demands of the concurrent load task, and hence reveal the underlying effect of concurrent load predicted by EXIT.

2.3 Experiment 2

In Experiment 2, I implemented a response deadline for both tasks in both conditions. In addition, I also changed from a doubled-up design (Table 2.1) to a single set of cue-outcome mappings (Table 2.2) with single-word symptoms (e.g. *feverish*).

I expect that the time limit will reduce the opportunities to compensate for the effects of concurrent load. As a result, I predict that the IBRE (*BC* → *rare*) will be reduced in the concurrent load condition, relative to the control condition.

2.3.1 Method

The current experiment used a Bayesian Sequential Procedure, in which data is collected until pre-defined bounds of evidence for null or alternative models are reached (Rouder, 2014; Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019). This meant that for Experiment 2, I conducted analyses after every $10^t h$ participant, starting when the sample size per group exceeded that of Experiment 1. This was repeated until all Bayes Factors were either below 0.3 or above 3.

Participants

Participants were undergraduate students at Plymouth University, who completed the experiment for course credit. Participants were randomly allocated to each condition, but I biased the counting so that there would be more participants in the concurrent load condition. This was because I anticipated higher numbers of exclusions in that group, due to the combination of concurrent load and time pressure. 160 participants took part in the experiment ($M_{age} = 20.69$), ranging between ages 19 and 59). 100 participants were allocated to the concurrent load condition and 60 participants to the control condition. Testing was conducted online.

Table 2.2: Abstract trial types during the training and test phase in Experiment 2.

Training (Relative Frequencies)	**Test (x20)**
$AB \rightarrow common_1$ (x 3)	A, B, C, AB, AC,
$AC \rightarrow rare_1$ (x 1)	BC

Apparatus and Materials

The experiment was implemented in Javascript with JsPsych (De Leeuw, 2015a). The words used were *dizzy*, *feverish* and *nauseous*; these words were randomly allocated to abstract cues shown in Figure 2.2. Words in compound cues appeared on separate lines one below the other, at font size 60px. Their position was counterbalanced.

Procedure

The digit and categorization tasks were largely the same as in Experiment 1, but with a response deadline of 5 seconds. In addition to the imposition of time pressure, the following further changes were made. Given that I only had a single set of cue-outcome mappings, I reduced training blocks to 8 trials. The common disease was presented six times and the rare disease was presented twice. Participants were trained either until they had completed 5 blocks, or until they achieved a criterion of two sequential errorless blocks, whichever came first. The test phase of Experiment 2 consisted of 120 trials, in which each test stimulus was presented 20 times.

Analysis

I applied the same thresholds for the Bayes Factor as before. In Experiment 2, there are just two diseases; one common and one rare. Hence, necessarily P(common) and P(rare) sum to one. Thus,

in this experiment, I used a one-sample Bayesian test comparing the probabilities of responding 'rare' on (*BC*) trials to random responding, ($\mu = 0.5$). A conclusive Bayes Factor and a mean group-level P(rare) above 0.5 indicates a rare preference. To test the effects of concurrent load, I directly compared the distribution of rare response probabilities of the two conditions, using a within-subjects Bayesian t-test.

2.3.2 Results and Discussion

I excluded participants who made more than one mistake at the final (5^{th}) block. In addition, I excluded a participant who had a high number of timeouts (25%) in the test phase. This meant that the final analysis included 52 people in the concurrent load condition and 53 people in the control condition. Figure 2.1 shows the probability of responding with rare for each test item in both conditions. In the control condition, I found overwhelming evidence for the presence of an IBRE, ($BF_{10} = 2,209$). In the concurrent condition, I found strong evidence for the absence of an IBRE, ($BF_{10} = 0.19$). Crucially, there was also clear evidence for the difference of choice proportions on BC between conditions, ($BF_{10} = 10.38$). Thus, after implementing a response deadline both in the digit and categorization task, and simplifying the overall experimental design, I observed that concurrent load reduced the size of the inverse base-rate effect This provides evidence in support of EXIT's striking and counter-intuitive prediction.

2.4 Chapter Discussion

In the inverse base-rate effect (IBRE), people learn that one combination of symptoms (*AB*) predicts a common disease, while an overlapping pair of symptoms (*AC*) predict a rare disease. When subsequently asked about symptom *A* alone, they predict the common disease, but when asked about *BC*, they predict the rare disease. This latter response is in opposition to the underlying base-rates of the two diseases and is, thus, arguably an irrational generalization. The IBRE has been the subject of formal mathematical modeling, most notably by the EXIT model (Kruschke, 2001b). This model makes the striking prediction that concurrent load during training should reduce the size of the IBRE. In other words, it predicts that distraction during learning should reduce irrational generalization in this procedure. In the current experiments, I found evidence in support of this prediction, but only where people were also under some time pressure; this time pressure presumably reduced their ability to compensate for the effects of concurrent load.

In summary, the current work leads to the conclusion that irrationality can sometimes be reduced when you are distracted. Although surprising, this conclusion is consistent with several other recently reported cases in which researchers observed better performance under increased load. For example, Borragán, Slama, Destrebecqz, and Peigneux (2016) observed a performance improvement in procedural learning when completed under concurrent load. Smalle, Muylle, Duyck, and Szmalec (2021) and Smalle, Daikoku, Szmalec, Duyck, and Möttönen (2022) similarly observed an improvement in phoneme and word learning. Overall, it can be argued that cognitive depletion improves some aspects of performance. In my case, it pushed people towards a more rational generalization. Irrationality is therefore not always due to low-level processes; sometimes it is the result of effortful high-level cognitive processes. There is some evidence for this from animal cognition; highlighting (a phenomenon closely related to the IBRE) is not observed in baboons (Fagot, Kruschke, Dépy, & Vauclair, 1998). Thus, it might be the case that the IBRE is uniquely human. Indeed, it may be limited to adult humans, as children have been reported to not exhibit the IBRE (Winman et al., 2005). Thus, the type of irrationality observed in the IBRE might arise only after higher cognitive functions are sufficiently developed. The ability to think more deeply seems to sometimes lead to increased irrationality.

One possible interpretation of all these results is that higher cognitive mechanisms, such as executive cognitive control, are detrimental in some tasks (Borragán et al., 2016; Smalle et al., 2022, 2021). Recently, Tsetsos et al. (2016) also put forth the argument that choice irrationality arises from purposeful and effortful thinking, unencumbered by any bottleneck on information processing. Our, and arguably the general, approach here is a multi-process account. In the framework of EXIT, the two processes are attentional learning and cue-outcome learning. Concurrent load causes the system to disengage attentional learning processes to accommodate the increased task demands.

2.5 Chapter Conclusion

In this chapter, I revisited a surprising prediction of a formal mathematical model of learning, EXIT (Kruschke, 2001b). The prediction is that, in the inverse base-rate effect, irrationality is reduced when task demands are increased; specifically when a concurrent load is imposed. I confirmed this prediction, but only in the case where people were also under time pressure. Overall, I found clear evidence for irrational generalization only under conditions where participants were not pressured by time constraints nor additional tasks. Thus, surprisingly, a *lack* of pressure is

sometimes detrimental to rational generalization.

Open Science

Experiment 1 can be found on osf.io/u8kcb/ and github.com/lenarddome/ply086-concurrent-ibre. Experiment 2 can be found similarly on osf.io/ng8kj/ and github.com/lenarddome/ply227-between-concurrent-minimal-ibre. Model simulations ar archived both on osf.io/vzhfk/ and github.com/lenarddome/ply228-exit-ibre-concurrent.

Chapter 3

Removing Error-Driven Componenents from the Inverse Base-Rate Effect

3.1 Introduction

In Chapter 2, I saw that the IBRE is sensitive to interference and that this observation corresponds with what EXIT predicts. One possible explanation for this reduction is a multi-process account, which is consistent with the internal operations of EXIT. Cue-outcome and attentional learning operate together but could engage differentially subject to task demands. This result also indicated that irrationality is effortful. The ability to think deeply leads to increased irrationality, which seems to be a prerequisite for the IBRE to arise. In this chapter, I focus less on models' predictions and learning mechanisms. Now I shift towards investigating the main assumptions underlying their success – error-driven attention. In order to do this, I will need to remove components from the experimental design that conceptually underpin certain components of the model. Simply put, I will modify the problem structure to be dissimilar to what the models are comfortable with as of today. Compared to Chapter 2, where I focused on attentional learning mechanisms, here I intend to fundamentally change the domain of problem structures where the IBRE was previously observed.

3.1.1 The underlying assumption of error-driven theories of the IBRE

All the most prominent theories of the IBRE involve an attentional mechanism that drives both learning and responding. These theories are also 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). As previously discussed in Chapter 1, 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, $AB \rightarrow common$, and C belongs to rare, $C \rightarrow rare$ (Kruschke, 2001a).

3.1.2 Current Chapter

In this Chapter, I intend to test this basic assumption of these theories by modifying the problem structure. In the following two experiments, I will gradually remove components from the design traditionally associated with prediction error. Our overarching goal is to investigate whether IBRE persists, even if I experimentally remove a crucial assumption of already existing accounts. In my first attempt, I implemented the canonical IBRE design with the caveat that category labels are presented in unison with features. In my second attempt, I further removed 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.

3.1.3 Related Work

To my 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 et al. (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 I use an effect size of $d = 0.46$ observed by Inkster, Milton, et al. (2022) and an α of 0.05 with a non-directional alternative hypothesis, the experiment has 24% power[1]. Given this information, this pilot experiment is underpowered. There are also no details about the procedure of this experiment. Therefore, I 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 instead matched the base-rate. These experiments give 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 et al. (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 (2005b)). 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 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 is driven by prediction error.

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

3.2 Experiment 3

Below, I detail my first attempt to test whether I 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 3 will present category labels simultaneously with their respective features. I retain the sequential property of the experiment, which means that participants learn about feature and category relationships on a trial-by-trial basis. I substantially simplified my implementation by removing the doubled-up design and reducing the number of test items to 6.

3.2.1 Method

Participants

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

Apparatus

The experiment was programmed in JsPsych (De Leeuw, 2015b) 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[2]. I 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.

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

Procedure

Table 3.1 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 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 3.1. 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 3.1: Abstract design of Experiment 3 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 for the presence of the IBRE, I calculated a Bayes Factor for a one-sample design. I calculate the probability of responding with the rare label on the critical BC test item, $P(rare|BC)$, for each participant. Then I tested this distribution of probabilities against the null, $mu = 0.5$, which denoted random responding. If the Bayes Factor fell below 1/3, I concluded

that participants' responses are not different from random responding. If the Bayes Factor fell above 3, I concluded that participants' responses reliably differ from null. If the mean probability of $P(rare|BC)$ is higher than 0.5, I conclude that I observed the IBRE. Values lower than 0.5 would indicate base-rate following. I used the method implemented in the BayesFactor R package (Morey & Rouder, 2022).

Exclusion

To match performance with the predictive learning implementations of the IBRE, I 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. I calculated the Bayes Factor for binomial proportions via the method implemented in BayesFactor R package (Morey & Rouder, 2022).

3.2.2 Results and Discussion

After exclusion, 125 participants made it into my main analysis. In summary, the qualitative pattern in my results corresponds to the base result of the IBRE. Table 3.2 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 $[0.63, 0.73]$, $\text{BF}_{10} = 2.45 \times 10^7$. 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 $[0.62, 0.72]$, $\text{BF}_{10} = 1.11 \times 10^7$. This gives me a sufficient amount of evidence to conclude that I have observed the IBRE.

Table 3.2: Group-level mean probabilities for each stimulus presented during the test phase in Experiment 3 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

Thus the current study strongly confirms that the IBRE can be observed in an observational proce-

dure. 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 4, I adress this by removing any design component that makes it clear to participants what the category label is. And I also use stimuli that reduces the chance of people assuming any causal relationship between its features.

3.3 Experiment 4

In this experiment, I implemented the IBRE in a way similar to cued-recall tasks. In this implementation, features were selected to be solid black geometric shapes and category labels were treated as features The task asked participants to memorize the arrangement of these shapes. On each trial, I 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.

3.3.1 Method

Participants

I recruited 171 undergraduate students who completed the experiment for partial course credit. Recruitment was done via the SONA recruitment system.

Figure 3.1: Simple geometric shapes used as stimuli in Experiment 4.

Table 3.3: Abstract design of Experiment 4 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

Stimuli

Stimuli were common solid geometric shapes, shown in Figure 3.1. Common and rare category labels were turned into features X and Y respectively. For each participant, each shape was randomly allocated to one of the abstract features shown in Table 3.3.

Procedure

Table 3.3 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 3, 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 3, 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 3.

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 3, 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

I applied the same analysis and exclusion methods as in Experiment 3.

3.3.2 Results and Discussion

After exclusion, 86 participants made it into my analysis. The group-level mean probabilities are shown in Table 3.4. The results are a qualitative and ordinal match to Experiment 3. Participants showed a clear common preference for stimuli A, $M_A = 0.78$, 95% HDI $[0.73, 0.83]$, $\text{BF}_{10} = 5.37 \times 10^{13}$.

Table 3.4: Group-level mean probabilities for each stimulus presented during the test phase in Experiment 4 after exclusion.

	$P(common)$	$P(rare)$
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 $[0.67, 0.79]$, $\text{BF}_{10} = 8.12 \times 10^8$. This gives me a sufficient amount of evidence to conclude that I have observed the IBRE.

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

3.4 Chapter Discussion

In this study, I 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 my first experiment, I 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, I 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 my second experiment, I 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. I still observed the IBRE.

The two experiments together suggest that the necessary conditions to observe the IBRE are fewer than previously established. In Experiment 4, 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 my 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 attentiontal 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 pinpointing the networks that uniquely underlie this rare preference but are independent of task demands.

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

3.4.1 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). $BC \rightarrow rare$ 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 I 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 $BC \rightarrow rare$ 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, $B \rightarrow common$, participants use the most similar rule applicable to BC from a "guessing set", $C \rightarrow rare$. This results in the $BC \rightarrow rare$ 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 $BC \rightarrow rare$ 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).

3.5 Chapter Conclusion

Across two experiments, I investigated whether the demonstration of the IBRE requires the prediction-and-feedback components of the standard experimental procedure. In Experiment 3, I 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 4 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. So, while in Chapter 2, I confirmed the prediction of the most complete class of explanations, here I pushed the IBRE further beyond the scope of all its current formal explanations.

Open Science

I have made available the two experiments written in javascript, the analysis code, the raw data, and all other supplementary materials both on the Open Science Framework and GitHub. Experiment 3 is shared on osf.io/auwvt/, and github.com/lenarddome/ply216-observational-ibre. Experiment 4 is similarly shared on osf.io/2tmc4/ and github.com/lenarddome/ply222-non-causal-ibre.

Chapter 4

Heterogeneity in the Inverse Base-Rate Effect

In Chapter 2 and Chapter 3, I arrived at a crossroads. I have confirmed a counter-intuitive prediction of the attentional explanations of the IBRE, but also showed how this irrational generalization effect does not depend on prediction error - the main driving force behind almost all of the best explanations. This means that the IBRE and associated phenomena have no formal explanation that could accommodate the complete domain of problem structures. So, I decided to take a step back and find the relatively most adequate model amongst the proposed explanations of the canonical and simplest version of the IBRE. In addition to the simple pursuit of group-level accommodation, I also decided to incorporate heterogeneity into the model benchmark. Human heterogeneity is not yet explored in the inverse base-rate effect. Model heterogeneity is also yet to be explored. This could give an indication of what starting point we should pick for approaching the current state of affairs from a modelling perspective.

This chapter contains work that is distinctly different from the first two strains of investigations. Here, I develop a general framework for model evaluations to solve a problem relating to the IBRE. To this effect, I will also discuss the framework's implications in addition to its implications for the IBRE.

Model Comparison Approach

The behaviour of formal psychological models[1] extends beyond their best-fitting parameter sets. Suppose I optimize model behaviour to reduce the discrepancy between the already observed empirical data patterns and the model outputs. In that case, the resulting model success will disregard the range of behaviours the model can produce. Most often, minimizing this discrepancy

[1] A *formal psychological model*, as defined here, is a mathematically-specified theory that is implemented as a computable algorithm for the purpose of simulating some aspect of human (or animal) behaviour.

underlies what is considered to be model adequacy. It is standard practice for psychologists to assess the adequacy of formal models by the extent to which they can accommodate an already-observed data pattern — a goodness-of-fit approach. This approach embodies the belief that the model with the best fit to the already-observed data is the one that best approximates the cognitive mechanisms at play in producing the behaviour, a belief that has been questioned before (Pitt, Kim, Navarro, & Myung, 2006). However, as Roberts and Pashler (2000) have previously argued, three aspects of model adequacy are not well addressed by a goodness-of-fit approach, at least as typically applied:

1. **Prediction**. Making predictions is sometimes seen as important to the evaluation[2] of scientific theories (Lakatos, 1976), but a good fit does not tell us what the theory predicts. One of the advantages of formal models is that they can make unambiguous and specific predictions that one can directly compare with data (Wills & Pothos, 2012). It can be useful to know what empirical, as-yet unobserved, predictions a theory makes — when I articulate a theory, I often want to know the observable consequences of that theory.

2. **Heterogeneity**. It's generally accepted that different people may do different things in the same experiment. It seems important to know the set of things people do and the subset of those things that a model accommodates. An exclusive focus on a good fit to group-level data risks building a theory of behaviours that no (or only a few) individuals in the group exhibit.

3. ***a priori* likelihood**. A focus on goodness-of-fit neglects the *a priori* likelihood that the theory will fit. If a model can accommodate literally any pattern of data, observed or unobserved, then its fit to any particular observed data pattern is largely meaningless.

In what follows, I propose and apply an alternative to the goodness-of-fit approach to model adequacy, an approach that addresses the issues raised by Roberts and Pashler. At the heart of my proposal is Robert and Pashler's second, fairly uncontroversial, assumption that both human and model behaviour is sometimes heterogeneous.

[2]We'd argue that Lakatosian approaches to the philosophy of science emphasize the role of predictions in evaluating theories, rather than developing them, which we'd argue is a creative endeavor in which "anything goes" (Feyerabend, 1975). However, expanding further on this aspect of the philosophy of science falls outside the scope of this thesis.

Heterogeneity

It is sometimes the case that different subgroups of people behave differently in the same experiment, and in ways that lead to the whole-group average being unrepresentative of some or all of the individuals that comprise it. For example, it is said that people either love or hate the taste of Marmite (a foodstuff). If true, a group-level rating of liking of 3 on a 5-point scale misrepresents this state of affairs. Although the existence of behavioural heterogeneity has long been appreciated (Merrell, 1931; Sidman, 1952; Blyth, 1972; Estes, 1956; Siegler, 1987; Ashby et al., 1994), it remains relevant today across numerous areas of inquiry. For example, heterogeneity of behaviour has recently been reported in spatial navigation, where strategies vary substantially between individuals (He et al., 2022). It has also been reported in studies of stimulus generalization, where group-level generalization gradients appear to be an aggregate of at least two distinct generalization types — linear and peaked (Lee et al., 2018). Similarly, in category learning, some individuals seem to apply rule-based strategies, while others apply similarity-based strategies (Shanks & Darby, 1998; Wills et al., 2011; Nosofsky & Hu, 2022).

In other words, the number of distinct subgroups of behaviour observed across individuals within a single experiment can be greater than one. Where this is the case, I argue that theories should be assessed against each of those distinct subgroups, rather than against the whole-group average. This is because, in such cases, whole-group averages obscure the heterogeneity of human behaviour, and may in some cases represent the behaviour of no (or very few) individuals.

A formal model can also be heterogeneous in its behaviour. This is because formal models often include a set of variable parameters that affect their operation. Varying these parameters can result in a range of different outputs for any given experiment. It can sometimes be convenient to think of variable model parameters as existing in a *parameter space*. For example, if a model has two variable parameters, then any pair of values for those parameters can be expressed as a point in a two-dimensional space. In that parameter space, different points can sometimes produce different model outputs.

Proposal

I propose that, in cases where human or model behaviour is heterogeneous, *the adequacy of a formal model is related to the extent to which it exhibits a similar range of behaviours to the humans it models.* This proposal differs from some other, more typically employed, methods of

assessing model adequacy, such as determining whether there is a single set of model parameters that result in the model closely approximating the whole-group average.

In order to express this proposal more formally, as I will do in the next section, I must first define some terms:

Accommodation When a model can reproduce a data pattern in at least one point of its parameter space, I say that the model *accommodates* the data pattern. What I describe as accommodation is sometimes described elsewhere as a *good fit* or a *model prediction*. However, accommodation is not prediction in the sense I mean it here (see below), and the concept of a fit being 'good' seems to largely neglect the question of whether there was any result observed or otherwise that the model could not have accommodated. Hence, I favor the more neutral term *accommodation*.

Prediction A prediction, for the purposes of the current proposal, is a data pattern that a model says should be observable in a particular experimental procedure, but that has not yet been observed. If that data pattern is later observed, it ceases to be a prediction and becomes an accommodated phenomenon (see above).

Types of inadequacy One broadly-accepted definition of a model inadequacy is a failure to accommodate an observed data pattern – the model is shown to be inadequate through a deficit of accommodation. However, there is another way in which a model can be inadequate. A model that predicts that literally anything can happen can never fail to accommodate an observed result. Such an overly-flexible model is inadequate through a lack of specificity of prediction. A model that makes predictions that turn out to be correct is sometimes considered to be a good or useful model. However, if a model makes a successful prediction merely because there is nothing it cannot accommodate, this does not seem to be a good model.

4.1 *g*-distance: A measure of model adequacy

In this section, I describe a formal measure of model adequacy, which is based on the concept of overlapping human and model heterogeneity previously outlined. Computation of this measure, which I call *g*-distance, proceeds in a series of steps, which are described in more detail in the sub-sections that follow.

First, model behaviour, which is typically continuously variable across changes in its parameters, is discretized at some level of granularity. Individual human behaviour is then discretized at the same level of granularity.

Second, all the different, discretized, behaviours a model can produce throughout its parameter space are derived through a process of *parameter space partitioning* (Pitt et al., 2006). A corresponding list is also compiled for all the discretized human behaviours so far observed in the procedure being modeled.

Third, I determine two metrics. The first, *accommodation* (alpha) is the proportion of behaviours observed in humans that are also produced by the model. The second, *prediction* (beta), is the number of behaviours produced by the model that have not been observed in humans, expressed as a proportion of all technically possible but as-yet-unobserved human behaviours.

Finally, these two metrics are combined into a single measure, *g*-distance, by conceptualizing model adequacy as the inverse of distance in space from the ideal model. The space considered is two-dimensional, with accommodation and prediction as the two axes. By default, accommodation and prediction are equally weighted, but a variety of positions on the relative importance of these two measures can be considered by weighting the two dimensions unequally. The choice of the letter *g* in "*g*-distance" is somewhat arbitrary (as is the choice of the letter *r* for a correlation coefficient, for example). However, as a mnemonic device, one might conceptualize 'g' as standing for 'ground truth', i.e. what is known about human behaviour in this procedure under conditions of perfect information. A model with a *g*-distance of zero produces exactly the same range of behaviours as humans do, with humans providing the ground-truth for the model.

4.1.1 Discretization

Typically, formal models in psychology produce output that varies continuously as the model parameters change. For example, a model might output the probability that an individual makes a particular response. This probability output can take any value between 0 and 1, and the output changes continuously as a parameter (e.g. response bias) is changed. Because the output probability is continuously variable, a model can in theory produce infinitely many outputs. And, among those infinite outputs, there would be an infinite number of results empirically indistinguishable from each other in humans at any given sample size. As Pitt et al. (2006) have shown, one effective solution to this problem is to discretize the model outputs into a set of patterns. This reduces the potentially infinite set of model outputs into a countable set of patterns.

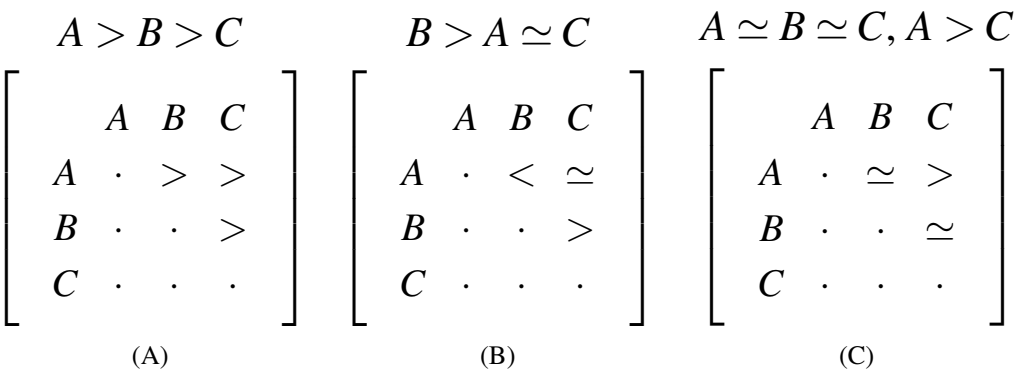


Figure 4.1: Three examples of the 19 possible patterns I could observe in a hypothetical experiment including three stimuli A, B, C. Each cell in a matrix compares two stimuli (row to column) on a dependent variable. Each cell can contain one of three values: < (smaller), > (larger), $\simeq$ (approximately equal).

The form of discretization used in my example application is ordinal discretization. Ordinal discretization is a natural fit to much of the informal theorizing that occurs in psychology, where one often builds experiments to test hypotheses that are expressed ordinally (e.g Group A will perform better than Group B). However, other methods of model discretization are possible, depending on the level of granularity in human data one wishes to investigate. I return to this point in Section 4.4.

To illustrate the concept of ordinal discretization more concretely, consider a simple experiment in contingency learning. I have stimulus A occurring with an outcome (A+). Later stimulus A and stimulus B occur in a compound with the same outcome (AB+). In addition, a new stimulus C also occurs but without an outcome (C-). After encountering all contingencies in this particular order, participants are asked to respond whether or not an outcome will occur if each stimulus, A, B, and C, are presented individually. A model of this experiment, with a given set of parameters, outputs the following probabilities that an individual will respond 'outcome occurs', A: .93, B: .47, C: .04. The ordinal discretization of this pattern is A > B > C.

While the pattern A > B > C is composed entirely of inequality relationships (<, >), an equality relationship is also possible. For example, the model may predict that the response probability for A and B are equal (A = B > C). Although the concept of equality is straightforward in a model, it is more nuanced for much human data, where measurement error can be substantial. Traditionally, experimental psychologists have focussed on whether or not there is substantial evidence for an inequality. In null-hypothesis significance testing, this is the difference between a significant and a non-significant result. In a Bayesian framing, Parameter Estimation (Kruschke & Meredith, 2021) defines the difference between equality and inequality in terms of an interval

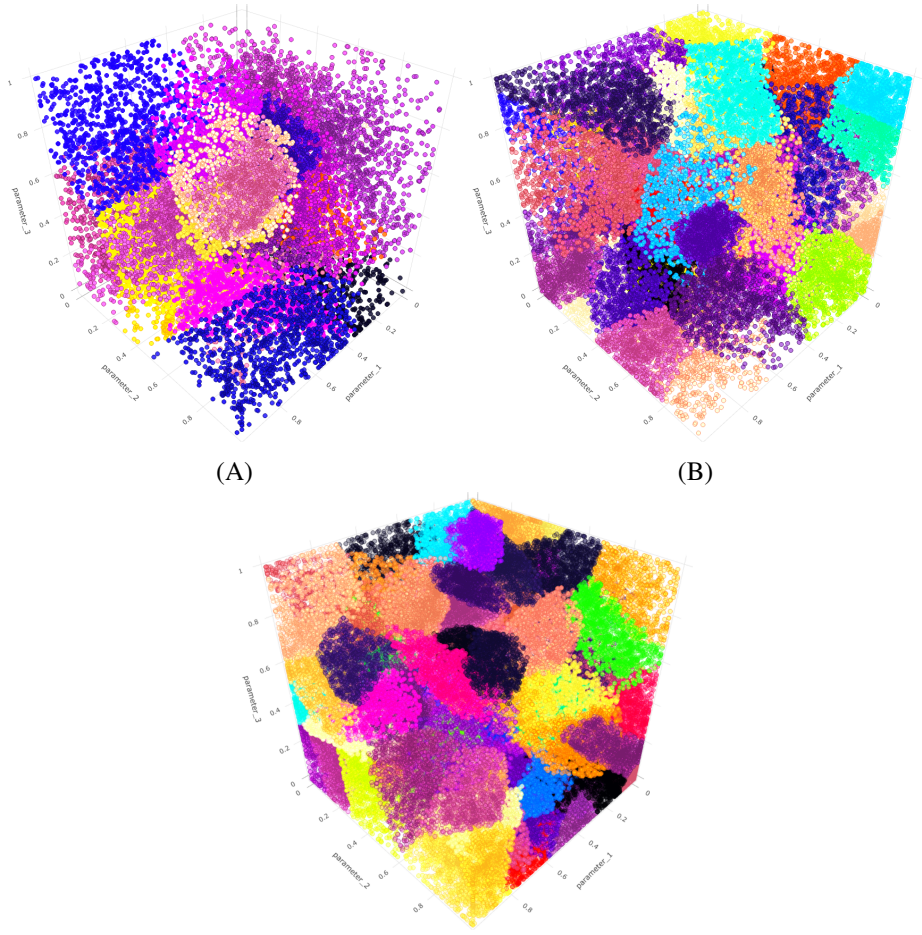
of beliefs. Inequalities are defined by how much of a difference I need to observe in order to treat that difference as reliable. In that sense, inequality can be considered to result from any difference that falls outside of a given interval of differences, while (approximate) equality is within that interval. Thus, in both traditional and Bayesian framing, one can consider ordinal patterns as being made up of inequalities (<, >) and approximate equalities ($\simeq$). Where the observed human data pattern includes an approximate equality, a model is considered to accommodate that approximate equality if the difference between its outputs is sufficiently small that it would be classified as an approximate equality if observed in a human. The process of model and human discretization is discussed in more technical detail in the Appendix B.3.2.

For an ordinal pattern with N components, there are necessarily M pair relationships in that pattern, where $M = N(N-1)/2$. Any N-component ordinal pattern can thus be fully represented by a set of M pair relationships. Here, I find it convenient to represent that set as a strict upper triangular matrix. Figure 4.1 shows three possible patterns for my illustrative experiment above, as matrices.

At first glance, one might assume that the number of ordinal patterns one could theoretically observe is 3^M, given that there are three types of pair relationship (<, >, $\simeq$) and M pairs (so, $3^3 = 27$ for my illustration). However, some of the patterns one can express in a triangular matrix are impossible (e.g. $A > B, B > C, C > A$). Where $M = 3$, the number of distinct possible patterns is 19. An algorithm for calculating the number of possible patterns for an N-component pattern is available in the R package *clobe* (Dome, 2023), discussed further in the Appendix B.3.3.

It is noteworthy that some of these matrices can also be represented by an ordered set of relationships, as done by Pitt et al. (2006). For example, the matrix in Figure 1B can be represented as $B > A \simeq C$, meaning that B is greater than A and C, which are approximately the same. Where $N > 3$, this ordered-set representation is more compact than a triangular matrix. For example, the pattern $A > B \simeq C \simeq D > E$ is more compact and arguably more readable than the 10-item matrix that would be needed in this case. However, not all triangular matrices can be represented by a single ordered set. For example, in Figure 1C, A is approximately equal to B, and B is approximately equal to C, which would lead to an ordered-set representation of $A \simeq B \simeq C$. Yet, A is greater than C, which is incompatible with this form of representation. For this reason, I recommend that the outcome of ordinal discretization is formally represented in a matrix rather than in an ordered-set form. However, for the subset of patterns that can be represented in ordered-set form, this provides a useful shorthand for the purposes of discussion and dissemination, and I use it in this thesis where appropriate.

4.1.2 Parameter Space Partitioning



(A) (B)

(C)

Figure 4.2: Parameter space partitioning for some "toy" three-parameter models. (A) A model producing 20 different patterns; (B) 50 different patterns; (C) 100 different patterns. Each point is a sample in parameter space; each color denotes a different model-behaviour pattern.

Following discretization, each pattern of behaviour that a model or a human emits is represented by a label, that label being a triangular matrix of the form shown in Figure 4.1. In order to compute g-distance, one must estimate the set of patterns produced by humans and the set of patterns produced by the model. The set of patterns observed in humans is estimated by large-sample data collection, as illustrated in a later section. The set of patterns produced by a model is discovered through parameter-space partitioning. Recall that, in the current approach, the goal is to discover *all* the things a model can do as its parameters vary, rather than find the single set of parameters that best accommodate a single, pre-defined pattern (as is done, for example, when one computes the goodness-of-fit to the group-level average).

Parameter space partitioning (PSP), looks for disjoint regions in the parameter space that elicit

specific discretized patterns of model behaviour. The details of the algorithm employed to achieve this are provided in Pitt et al. (2006), and in Appendix B.3.4. Figure 4.2 illustrates the outcome of parameter space partitioning for three toy models, each of which has three parameters that vary continuously between 0 and 1. In each case, the PSP algorithm correctly discovers all distinct patterns of behaviour the model can produce (20, 50, or 100 patterns, depending on the model). It also reveals the disjoint region or regions of the parameter space that produce that behaviour pattern. In the Figure, these regions are shown as clouds of dots, where the color of the dot denotes the pattern observed, and the area approximated by the cloud of same-color dots is the region or regions of parameter space that produces that behaviour. In the current application, it is the discovery of patterns that is crucial rather than the location or size of the regions. The possibility of using volumetric measures is considered in the Chapter Discussion.

While the preceding description of PSP obscures some technical detail of how the outcome is achieved, the goal seems, and is, straightforward – to enumerate all the different things a model can do. It is, however, much more computationally expensive on average to enumerate all the behaviours of a model, than to assess its ability to accommodate a single pattern. It is perhaps for this reason that PSP, despite its conceptual simplicity, has so far not been much used in psychology, beyond the examples used in its original introduction (Pitt et al., 2006). As we'll illustrate in a later section, computing power is now such that, along with the efficiently-coded models and PSP algorithms now available (Wills et al., 2022; Dome & Wills, 2023), parameter-space partitioning can be achieved in manageable time on consumer-grade hardware. This situation will likely improve further over the next several years.

4.1.3 Calculating g-distance

Once one has enumerated both human and model behaviours, g-distance can be calculated. The process of calculation, informally described earlier in this section, is set out more formally below. This allows more precise, compact expression while illustrating a number of concepts we'll make use of later. Readers unfamiliar with set notation may wish to consult the Glossary.

Consider the sets in Figure 4.3. H is the set of data patterns that humans have been observed to exhibit, under some defined conditions (e.g. in a particular experiment). M is the set of data patterns that the model can produce, under those conditions. Both are subsets of the universal set (U) of all conceivable data patterns - both observed and unobserved, both producible by the model

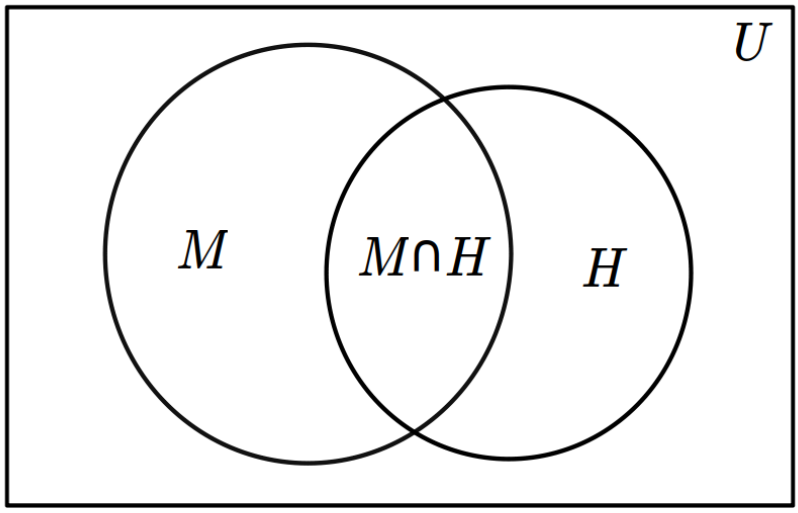


Figure 4.3: The sets employed in the calculation of g-distance.

and not producible by the model.

In the case of perfect information, where both H and M are completely known, a perfect formal theory would accommodate everything in H, but predict nothing, as any prediction would be, by definition, incorrect. Thus, where H is completely known, and my model is perfect, ($M' \cap H = \emptyset$) (complete accommodation) and ($H^{'} \cap M = \emptyset$) (no predictions). Hence, in the case of perfect information, model adequacy can be considered as a function of how close ($H^{'} \cap M$) and ($M' \cap H$) are to being empty sets.

Quantifying this, I express a model's *sufficiency of accommodation* (α) as the number of human data patterns accommodated by the model, as a fraction of the total number of observed human data patterns,

$$\alpha = \frac{|M \cap H|}{|H|} \tag{4.1}$$

I further express a model's heterogeneity in terms of its *breadth of prediction* (β), which is the number of unobserved human data patterns the model predicts, expressed as a proportion of the total number of unobserved human data patterns.

$$\beta = \frac{|M \cap H^{'}|}{|H^{'}|} \tag{4.2}$$

Under the conditions of complete information earlier defined, a perfect model has an α of one and a β of zero. I call this the *PAS* (Perfect Accommodation and Specificity) point, specificity being defined as $1 - \beta$.

From these definitions, I operationalize model inadequacy as the Euclidean distance from the PAS point in a two-dimensional Euclidean space

$$g = \sqrt{\frac{(1-\alpha)^2 + \beta^2}{2}}, \tag{4.3}$$

In the case of complete information, g is a measure of the distance of the model from the 'ground truth' of the observations. I refer to this as *g-distance* (Ground Truth Distance).

Equation 4.3 attributes equal weight to accommodation and specificity, which seems acceptable as a default position, but this is not a strong theoretical commitment on my part. I anticipate, for example, that some may wish to weigh accommodation more heavily than specificity. Thus, I express *g-distance* as

$$g = \sqrt{w_\alpha(1-\alpha)^2 + (1-w_\alpha)\beta^2}, \tag{4.4}$$

where $0 < w_\alpha < 1$. Operationalizing g-distance in this way also means that the statement Model A is more adequate than Model B ($g_A < g_B$) can be conditionalized for an interval of beliefs about the relative importance of accommodation and specificity (e.g. $g_A < g_B$ where $1 < w_\alpha < 0.25$).

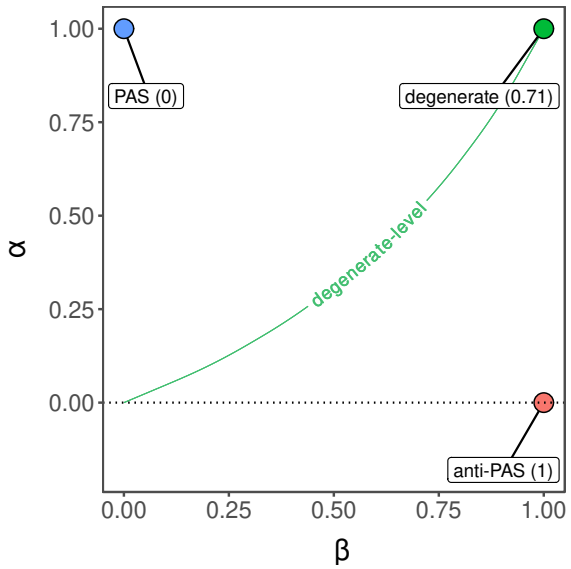


Figure 4.4: The figure shows the positions of a PAS model, a degenerate model, and an anti-PAS model as a reference. The green line represents an illustrative threshold.

The *g-distance* metric is designed primarily to *compare* the adequacy of two or more candidate models. For example, where $g_A < g_B$, Model A is considered to be the more adequate model under this metric. However, in addition to comparative adequacy, it is possible to define some threshold points on the g-distance metric. Consider three hypothetical models that are at the following threshold points – the PAS point, the degenerate point, and the anti-PAS point. These points are depicted in the on Figure 4.4. To make the interpretation straightforward, I make the assumption that I know all empirical information there is to know about this phenomenon and I am testing different theories I built to explain it.

The **PAS point**, previously described, is the location of an ideal model under these conditions. It has an alpha of 1 and a beta of 0, and hence a g of 0.

In contrast, the **anti-PAS point** is the location of the worst possible model. It accommodates nothing that is observed and predicts everything that is not observed. It has a g of 1.

A model at the **degenerate point** is one that accommodates everything that is observed, but also predicts everything that isn't observed. It's degenerate in the sense that it assumes everything that can in theory be observed, is in fact observed. In this way, it is indistinguishable from an account that human behaviour is unlawful; a form of random model and one that, I suspect, few psychologists would accept as a good theory of behaviour. By this definition, a degenerate model has an α of 1 and a β of 1, and hence a g of $\sqrt{1-w_\alpha}$, which is $\sqrt{0.5}$ (approx. 0.707) when accommodation and specificity are weighted equally (i.e. $w_\alpha = .5$).

I further argue that any model with a g-distance of $\sqrt{1-w_\alpha}$ or higher is a poor model, as its

adequacy is no better than that of a degenerate model. This seems uncontroversial but leads to the conclusion that there are models which, while not degenerate, are also poor models on this metric. I describe these as *degenerate-level* models, and suggest that being above a degenerate level of adequacy should be considered a minimum threshold for model adequacy[3]. In the interval below $\sqrt{1-w_\alpha}$ down to zero, *g*-distance is intended to be largely comparative, rather than having further specified thresholds.

4.2 Applying *g*-distance to the inverse base-rate effect

In previous sections, I set out the rationale and implementation of the g-distance metric of relative model adequacy. In the current section, I apply the metric to five models of the inverse base-rate effect (IBRE) (Medin & Edelson, 1988).

All the model implementations and analysis tools used in this section are available as open-source software, released as *R* packages (R Core Team, 2023). Model implementations are available in the *catlearn* package (Wills et al., 2022). The parameter space partitioning routine is implemented in the *psp* package (Dome & Wills, 2023). Additional algorithms, including the g-distance equations, an inequality-matrix constructor, and an algorithm for deriving the universal set, are available in the *clobe* package (Dome, 2023). All but the *catlearn* package were developed during the course of the current project; a project that also led to the addition of five formal models to the *catlearn*. Simulation data were too large to store on GitHub, so they are only available on OSF.

The rest of this section comprises four parts. First, I introduce the Procedure. Second, I briefly describe the five models of the IBRE I will be comparing. Third, I describe and analyze a large-sample (N > 300) IBRE experiment I ran in order to provide high-quality data for model comparison. Fourth, I describe computing *g*-distance for these five models against my IBRE dataset.

4.2.1 Models candidates of the inverse base-rate effects

In what follows, I assess a total of five formal models of the IBRE against a large-sample experiment. These models fall into three sets of related models.

[3]Through re-arrangement of Equation 1, we see that for any value of α, a model has degenerate-level adequacy if β equals $\sqrt{1-(1-\alpha)^2}$, assuming alpha and beta are weighted equally. This function is not linear, and its particular non-linear form has some interesting consequences for model-improvement efforts. For example, if a model has low accommodation (below about .35), then one can increase the proportion of unobserved predictions (beta) more rapidly than the proportion of accommodations (alpha), without falling to degenerate-level accuracy. For models with higher levels of accommodation, the opposite is true. Thus, the importance of 'keeping a lid' on new unobserved predictions increases as a model becomes better at accommodation.

The first set of models are **attentional-associative** models, whose theoretical roots can be traced back to work by Mackintosh (1975). These models all explain the IBRE in a similar manner; they assume that people learn to redirect their attention to stimulus features in order to avoid errors. These models have already been discussed in Chapter 1.2, but it is worth presenting a brief recap before moving on to another class of models. In the context of IBRE research, the paradigm example of an attentional-associative model is the EXemplar-based attention to distinctive InpuT (EXIT) model (Kruschke, 2001b). Given its prominent status in the IBRE literature, it is thus hard to imagine any assessment of the relative adequacy of models of the IBRE without including EXIT. Nonetheless, EXIT is far from the only instantiation of error-driven attention. Paskewitz and Jones (2020) presented multiple simpler derivatives of EXIT. Thus, in addition to EXIT, I also consider two simpler models derived from EXIT by Paskewitz and Jones (2020) – the Neural Network with Competitive Attentional Gating (NNCAG) and Neural Network with Rapid Attentional Shifts (NNRAS) models. Each takes a subset of the processes in EXIT to produce a simpler, EXIT-like model. One might anticipate that, if EXIT is over-determined, one or both of these simpler models would accommodate the empirical results just as well as EXIT, and might at the same time make fewer unobserved predictions. This would lead to them being considered more adequate than EXIT on the g-distance metric. I were interested to see the extent to which this intuition was confirmed by analysis.

The second set of models I consider are **plain-associative** models. Like attentional-associative models, these assume that people learn by the formation of associations between the stimulus features and the category labels. However, unlike the attentional-associative models, there is no process of attentional reallocation. These models are largely considered to be poor accounts of the IBRE, being unable to capture the full set of known behavioural results in the IBRE procedure. I thought it might be informative to include a 'known poor' model in the set, for the purposes of comparison. Intuitively, calculating g-distance for a model broadly considered to be inadequate in a group-level goodness-of-fit approach might provide a useful lower bound against which other, potentially more adequate models, could be compared. For these purposes, a single, simple model is sufficient; I chose the Least-Mean-Square NETwork (LMSNET) by Gluck and Bower (1988).

The third set of models I consider are **dissimilarity-exemplar** models. The theoretical roots of these models can be traced back to exemplar-based theories of categorization, such as Nosofsky (1984), which assume that categorization proceeds by a process of computing the similarity of the current stimulus to stored representations of specific exemplars of each of the categories under consideration. Dissimilarity-exemplar models assume that, in addition to calculating the similarity

of a stimulus to examples of category X, one also calculates the dissimilarity of that stimulus to all other (non-X) categories under consideration. Like attentional-associative models, dissimilarity-exemplar models assume that there is a process by which different stimulus features can acquire different levels of attention. However, unlike attentional-associative models, there is no process specified by which these attentional weights are learned. Instead, the attention attracted by each stimulus feature is a free parameter of the model. In a typical model-fitting approach, these parameters are chosen in order to maximize goodness of fit to the group-level average. Intuitively, it seems like this approach might lead to many more unobserved predictions than the process-based approach taken by *attentional-associative* models. Again, I were interested to see the extent to which this intuition was confirmed or disconfirmed by analysis.

Although the preceding description of dissimilarity-exemplar models might be construed as criticism, dissimilarity-exemplar models should not be considered as *straw men*. Rather, they include a recently-published formal model, independently forwarded as a credible account of the IBRE, and the only formal model (other than EXIT) to be applied to the results of a neuroimaging study of the IBRE (O'Bryan et al., 2018) Thus, they are a class of formal account that, if excluded from consideration, would make my assessment of relative adequacy excessively partial. In what follows, I consider the version of the Dissimilarity Generalization Context Model used in O'Bryan et al. (2018)'s imaging study.

In summary, in the current article, I consider five distinct models that fall into three classes of account (including one 'known poor' class). These five models are described more fully in Appendix B.2.

4.2.2 Experiment 5

Although multiple published replications of the inverse base-rate effect exist (see the review of Don et al., 2021), which includes Experiment 1-4, none were suitable for the current application. This was because most IBRE experiments are relatively small-sample (N $\simeq$ 30). Although this demonstrably provides adequate power to detect the group-level results, it is unlikely to be sufficient to estimate human heterogeneity, assuming substantial heterogeneity exists. The question of appropriate sample size under human heterogeneity is not straightforward, as it depends strongly on assumptions about the extent of heterogeneity present, and on the frequency distribution of component patterns – information not available to us at design time. Thus, for this initial investigation, I simply collected data from as many participants as possible given the time and resources

at my disposal. This resulted in a sample of 354 participants, about an order of magnitude larger than the typical IBRE experiment.

The current experiment employed a simplified version of the canonical inverse base-rate paradigm. I decided to reduce the experiment to its most basic necessary conditions that I hypothesized would still give rise to the phenomenon: a single set of features and categories; a single shared, a single common predictive, and a single rare predictive cue; and a 3:1 common to rare ratio the canonical difference in base rate.

Table 4.1: Abstract design of the IBRE.

Training (Relative Frequencies)	**Test**
$AB \rightarrow 1$ (x 3)	A, B, C,
$AC \rightarrow 2$ (x 1)	AB, AC, BC, x 12

Participants

I collected data from 354 participants. This sample was collected via SONA (Systems, n.d.) and Prolific (Lange, Kühn, & Filevich, 2015). From SONA, I recruited 117 Psychology students studying at the University of Plymouth, who completed the experiment for course credit. After preliminary analysis based on an earlier version of my method, I decided to collect more data, given the large number of ordinal patterns I saw (Human Set) and I could have seen (Universal Set). I thus recruited a further 237 participants from Prolific, who received financial compensation of č2.50 for participation. The only restriction I set on Prolific was that participants must be able to speak English. This resulted in a Prolific participant pool spanning 24 countries across 5 continents. The age of participants ranged from 19 to 55, with a mean age of 28.6. The participant pool consisted of 112 people identifying as Male and 118 people identifying as Female, and 7 people with missing data (no demographic data was retained for the SONA sample). Further demographic information can be found in the online-available supplementary materials, see Open Science Statement at the end of this chapter. The experiment on average lasted no longer than 15 minutes.

Stimuli and apparatus

The abstract stimuli employed in the experiment are shown in Table 4.1. The physical stimuli were words describing symptoms: 'dizzy', 'feverish', and 'nauseous'. Each word had a 60pt font

size. These symptoms were randomly allocated to the abstract feature denoted by capital letters in Table 1. This was done for each participant when they opened the experiment in the web browser. Physical features, and symptoms, were mapped to disease categories, denoted by X and Y. Response keys corresponded with category names, X and Y. These categories were randomly allocated to the common and rare diseases at the beginning of the sessions. The order in which the physical features were presented on screen was counterbalanced across trials - for example, AC appeared equally often as AC and CA within each block. Words appeared above/below each other.

The experiment was written in JSPsych 6.1.0 (De Leeuw, 2015a) and deployed via JATOS 3 (Lange et al., 2015). Participants completed the experiment via a personal computer. Tablets and phones did not allow participants to respond and go beyond the welcome screen.

Procedure

In the training phase, participants were asked to learn the relationship between symptoms and diseases. The training items are shown in Table 1. On each trial, participants were presented with the stimuli for 5 seconds and were asked to categorize them into either Disease X or Disease Y; the stimuli were response terminated. During the training phase, each response was followed by feedback on whether the participant had made the correct or wrong response; the feedback was displayed for 1 second. There was then a 1-second inter-trial interval. Participants completed blocks of 8 trials with a 3:1 common-to-rare trial ratio - 6 common and 2 rare trial types per block. The trial order was randomized within each block. Participants were assessed against a learning criterion of 2 errorless blocks. If they completed two errorless blocks, they were transitioned immediately to the test phase. If the participant had not reached this criterion after five blocks, they were transitioned to the test phase anyway.

In the test phase, participants were asked to categorize new and old combinations of symptoms into Disease X or Y on a trial-by-trial basis, but without the corrective feedback. The test items are shown in Table 1. Participants completed 6 blocks of 12 test trials, encountering each test item twice in each block.

4.2.3 Group-level results

Figure 4.5 presents the results of my experiment in the traditional, group-level manner. The y-axis is the probability of responding that the stimulus belongs to the rare category. The experiment is two-alternative forced-choice, and hence a mean group-level probability exceeding .5 indicates

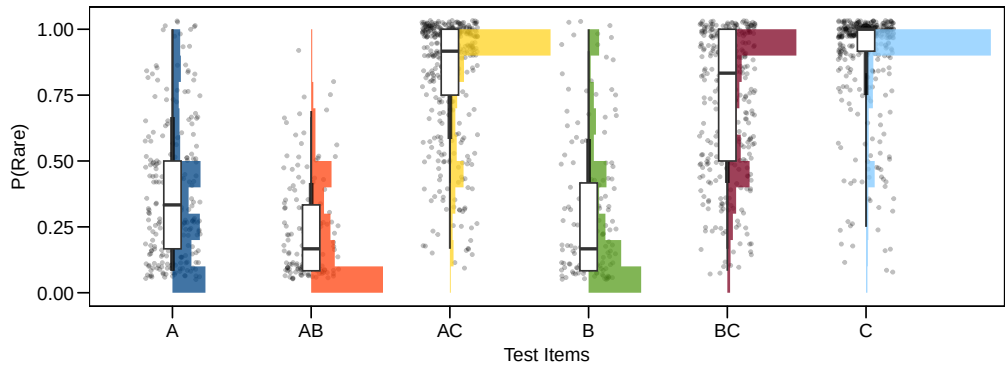


Figure 4.5: Probability to respond with the rare disease for each test item. Each dot is a participant. Vertical histograms show the number of values falling into that interval.

a preference to respond 'rare'. Thus, the IBRE is observed, as the mean probability of a 'rare' response exceeds .5. This is a conclusion that can be drawn with overwhelming levels of certainty at the group level, $M = 0.72$, 95% HDI $[0.69, 0.74]$, $\text{BF}_{10} = 7.29 \times 10^{34}$. The group-level common and rare preferences for the other cues are supported by yet-stronger evidence. For example, there is a common preference for A, $M = 0.75$, 95% HDI $[0.72, 0.78]$, $\text{BF}_{10} = 8.73 \times 10^{43}$. Although not my primary purpose in collecting these data, the current experiment is arguably the most compelling demonstration of the group-level IBRE to date.

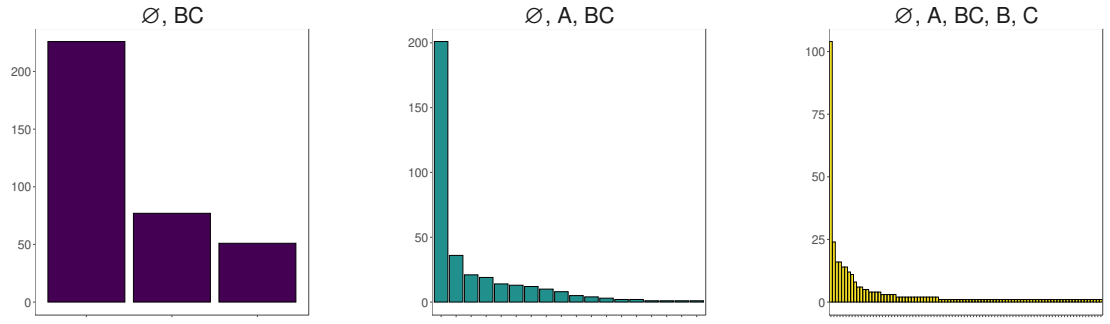
4.2.4 Individual-level results

Table 4.2: The top four ordinal patterns and the canonical group-level pattern (last row) with their corresponding frequency - the number of people who exhibited them.

Ordinal Pattern	Frequency
$A \simeq B < 0.5 < BC \simeq C$	104
$A < B < 0.5 < BC \simeq C$	24
$A < B < 0.5 < C < BC$	16
$A \simeq B < 0.5 < C < BC$	16
$B < A < 0.5 < BC < C$	3

As previously discussed (see Section 4.2.1), the current demonstration uses ordinal discretization to convert the potentially infinite set of model outputs into a countable set of patterns. In order to compare a set of model patterns with human data, the human data must also be discretized in a comparable manner; the algorithms I used for discretization are described in Appendix B.3.2.

In this demonstration, I chose to compare models to human data at three different levels of *gran-*



(A) (B) (C)

Figure 4.6: The frequency distribution of observed patterns at three levels of granularity. The y-axes show the counts of participants showing the ordinal pattern, while the x-axes enumerate the ordinal patterns. The x-axes are ordered by pattern frequency.

ularity. At the coarsest grain, I considered just one ordinal comparison - is the probability of responding 'rare' to BC greater than, less than, or approximately equal to, 0.5? Using 0.5 for the mid-point, the three possible patterns can be represented as $BC > 0.5$, $BC < 0.5$, and $BC \simeq 0.5$. When discretized at this level, all three possible patterns are observed in my human sample, although not with equal frequency, as illustrated in Figure 4.6A.

At the next, slightly finer, grain of discretization, I also included stimulus A. At this grain, the group-level pattern can be represented as $A > 0.5 > BC$, which captures the group-level result that there is a common-disease preference for A but a rare-disease preference for BC. Although perhaps not immediately obvious, it is the case that there are 19 possible three-stimulus ordinal patterns (the algorithm I used to calculate this is described in Appendix B.3.3.). Human participants produce all but one of the 19 patterns in the Universal set, with unequal frequency (Figure 4.6B).

At the finest grain I considered, I also added test stimuli B and C. At this level of granularity, the Universal set contains 2,131 patterns, while the Human set contains 90 patterns, with the now-characteristic decelerating-decay frequency distribution shown in Figure 4.6C. This is the first level of granularity at which the number of distinct patterns observed in humans is small relative to the Universal set. This is potentially useful; in particular for detecting situations in which the number of patterns produced by a model substantially exceeds the number observed in humans.[4]

[4]Although it would in principle be possible to consider yet finer grains (by, for example, including the AB and AC test items), two pragmatic issues meant I did not do so. First, the time parameter-space partitioning took to complete turned out to be, for the models under test, a rapidly accelerating function of the pattern granularity; going beyond the highest level of granularity considered above was beyond the computational resources available to us. Second, beyond the highest level of granularity considered above, the number of human-observed patterns approached the number of participants, meaning that all patterns were of approximately equal frequency. Given the decaying frequency distribution observed at previous granularity levels, I suspected that the flat frequency distribution observed at finer grains was an artifact of insufficient sample size. It was beyond the resources at my disposal to expand the sample size sufficiently to test this hypothesis.

Table 4.2 shows the four most commonly observed patterns at this finest grain of analysis, plus the group-level result. Perhaps the most immediately striking thing about these results is that less than 1% of the sample shows the group-level result. Looking closer, I also observe that all the top four patterns include ordinal relationships that were not present at the group level. These include components that are directly opposite to the group-level pattern and which, informally, seem like they might be hard for any of the models considered to accommodate. Of particular note are pattern components such as (in terms of the probability of responding 'common-disease') $A > B$ and $C > BC$. Under any of the modeling approaches considered, it is far from obvious how a cue associated with both outcomes (A) can have a higher P(common-disease) than a cue only ever associated with the common disease (B). Similarly, it seems difficult to explain within these models how a cue only ever associated with the rare outcome has a higher P(common-disease) than a cue compound where one component is associated with the rare disease and the other with the common disease (BC). In the following section, I will see to what extent this informal hypothesis of less-than-perfect accommodation by the models is confirmed by formal analysis.

4.2.5 Computing g-distance

Below I demonstrate that all five models under consideration are poor models of the IBRE experiment presented above when assessed by *g*-distance. In these calculations, α and β were equally weighted; the poor performance of all models meant that an exploration of unequal weights was not necessary in this case.

In order to calculate *g*-distance for the EXIT, NNCAG, NNRAS, and LMSNET models it is important to realize that these models are sensitive to the trial order during training. This means that, for a fixed set of parameters, they can produce more than one output pattern; the pattern they produce can depend on the order in which the training trials were presented. In the current experiment, each participant experienced a unique trial order (due to trial order randomization, which is standard practice in the field). Thus, for these four trial-order-sensitive models, the set of model outputs (M) revealed through parameter-space partitioning may be different for each participant. I addressed this by conducting a separate parameter-space partitioning and subsequent *g*-distance calculation for every participant's trial order. Under this approach, for any given participant, α is either 1 or 0 (the pattern produced by that participant is either produced by the model or it isn't), and β is potentially different for every participant (because different trial orders potentially produce different sets of output patterns in the model). The reported values for α and β in these

calculations are the averages across the individual-participant values.

The DGCM18 model, in contrast, is trial-order insensitive – for a given set of parameters, it produces the same set of output patterns, irrespective of trial order. It can nonetheless be used in the same method of calculation described above. For each participant, the pattern produced is either in DGCM18's set (M), or it isn't. Similarly, for each participant, there is a set of patterns that DGCM18 produces but the participant does not ($|M \cap H'|$). Thus, g-distance for DGCM18 can be calculated in the same averaged manner as in the other models.

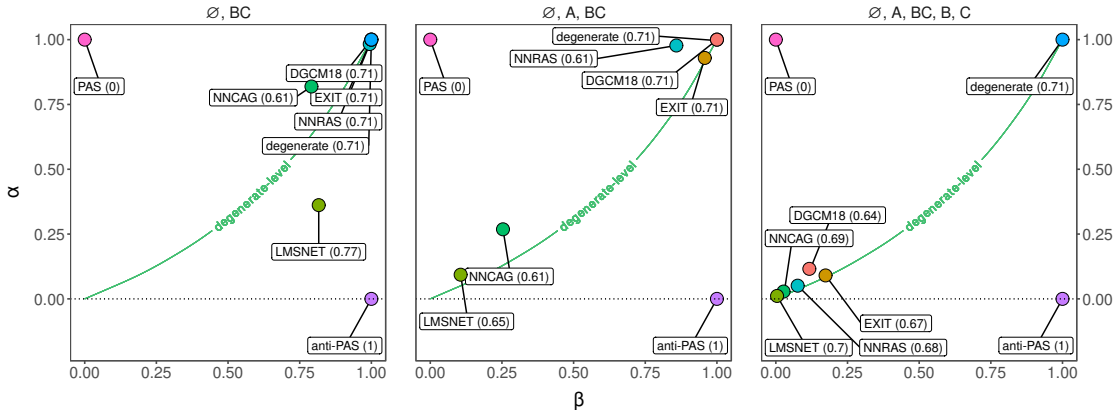


Figure 4.7: The three panels show all five models in a two-dimensional space, where α and β comprise the two axes of this space. Colored dots correspond to models. Labels show the name of the model and their corresponding g. Models with degenerate-level adequacy would appear somewhere on the green line.

Figure 4.7 shows the values for g, α, and β for all models at all three granularity levels. At the [0.5, BC] granularity, EXIT, DGCM18, and NNRAS are degenerate models, accommodating everything that is observed, and predicting everything that is not observed. LMSNET is worse than degenerate-level because its lower rate of unobserved predictions is more than offset by its much lower rate of accommodations. NNCAG is better than degenerate-level for complementary reasons; its lower accommodation rate is more than offset by its lower prediction rate. Thus, NNCAG is the most adequate model at this level of granularity. It is also the only model with better than degenerate-level performance. Nonetheless, while NNCAG accommodates 82% of participants' patterns, which seems quite good, it achieves this at a cost of predicting, on average, 79% of the patterns that could theoretically have been observed but weren't. So, while NNCAG is the most adequate model in this comparison, the other models set a pretty low bar.

At the next level of granularity [0.5, A, BC], DGCM18 remains a degenerate model, and EXIT also has degenerate-level adequacy. The other three models are all somewhat better, and all for basically the same reason. Each of the remaining models makes fewer unobserved predictions, but

at the cost of some capability to accommodate observed behaviours. This ranges from LMSNET which predicts almost nothing at the cost of accommodating almost nothing, to NNRAS which accommodates almost everything at the cost of predicting almost everything that could theoretically have been observed but wasn't.

At the finest grain, [0.5, A, B, BC, C], all models show poor accommodation, capturing between 1% and 12% of the paticipants' behaviours. However, this loss of accommodation, relative to coarser grains, does not reduce model performance as measured by g-distance. This is because, at this fine grain, the models also predict a small proportion of the large number of unobserved human behaviours. Overall, they all remain clustered close to degenerate-level performance, with DGCM18 now slightly better than the rest. Expressing this in terms of numbers rather than proportions may make the same point more starkly. For example, EXIT produced on average about 370 predictions for each trial order (and hence for each participant), and yet the behaviour actually emitted by the participant only appeared in that large set for 32 of the 354 participants.

As previously discussed Section 4.3.2, NNCAG and NNRAS are simplified versions of EXIT, with each simpler model taking a different subset of the component processes that make up EXIT. One might have expected this to result in lower flexibility in the simplified models than in EXIT, and this was indeed shown to be the case – at the finest grain of discretization, EXIT has a β more than five times larger than NNCAG for example. However, this reduction in flexibility was accompanied by a substantial reduction in accommodation, with the overall result that all three models turned out to be approximately equal in their (in)adequacy. Interestingly, the number of patterns produced by EXIT is much more affected by trial order than its simpler counterparts, as Figure 4.8 shows. Depending on the trial order, EXIT produced anywhere between 11 and over 600 patterns. The degree of variation is much lower for the other two models.

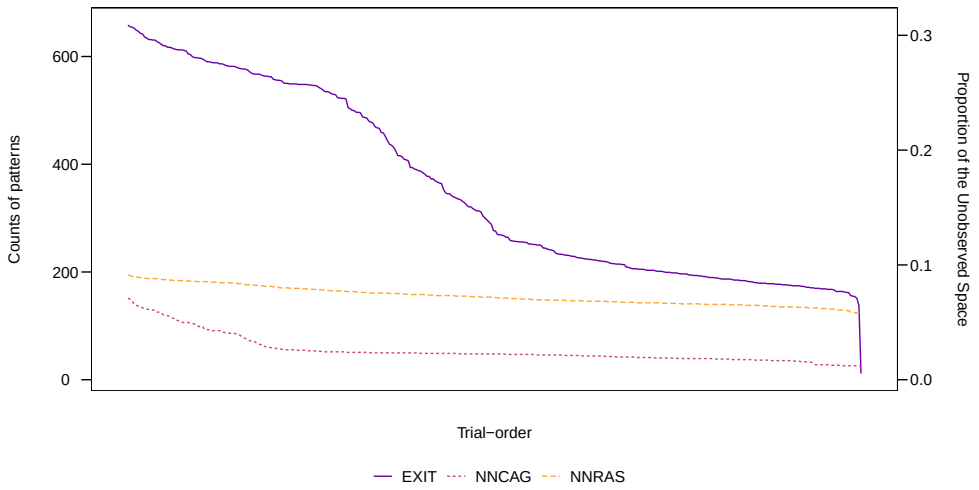


Figure 4.8: The counts of ordinal patterns (y-axis) three trial-order-sensitive models predict for each individual's trial order on the highest complexity with their corresponding trial order (x-axis). The x-axis is ordered by the number of patterns EXIT predicted.

In summary, at all three levels of granularity considered, all models were at, or close to, degenerate levels of adequacy. Of the five models I have examined, none could be described as adequate accounts of the inverse base-rate effect, where adequacy is indexed by g-distance.

4.3 Discussion

In this chapter, I introduced g-distance, a metric and framework for evaluating and comparing formal models. More specifically, g-distance evaluates the extent to which the range of behaviours produced by a model coincides with the range of behaviours observed in the task being modeled.

The development of g-distance was inspired by previous discussions of the tendency for theorists to neglect the likelihood of a model being a good fit to *any* conceivable set of data, rather than just the data that was actually observed (Roberts & Pashler, 2000; Gregg & Simon, 1967; Pitt & Myung, 2002). Using parameter space partitioning (Pitt et al., 2006), the g-distance approach uncovers the set of behaviours produced by a model and then asks two questions: (1) to what extent are the behaviours produced by humans in the set of model behaviours (accommodation, α), and (2) to what extent does the model produce behaviours I have not observed (prediction,

β). *g*-distance then combines α and β into a single measure, *g*, which represents the distance of the model from the ideal model under conditions of perfect information about human behaviour; under these conditions, the ideal model would accommodate all observed behaviours and produce no unobserved predictions (because, under perfect information, any unobserved predictions would by definition be wrong). The value of g-distance ranges from zero to one, where zero is the ideal model, one is the worst possible model, and $\sqrt{.5}$ (.71) is the point of degenerate-level adequacy; this means the point at which, on the *g*-distance metric, the model is exactly as adequate as an entirely random model.

Using the g-distance approach, I evaluated five formal models against a large-sample experiment on the inverse base-rate effect (Medin & Edelson, 1988). This demonstration turned out to have implications for several theories of this effect. Many of these models are designed to accommodate the group-leve results. Strikingly, all of the models I evaluated were close to the point of degenerate-level adequacy – in other words scarcely more adequate on the *g*-distance metric than a random model.

This poor performance is due to two, distinct, forms of inadequacy on the part of the models. First, it turns out that the group-level result observed in the IBRE paradigm is exhibited by less than 1% of individual participants. The group-level result appears to be an average of a set of distinct individual behaviours, with some of the most common seemingly hard to accommodate within any of the models considered. Thus, models such as EXIT, which were designed to accommodate the group-level result, fail to accommodate the majority of the individual behaviours observed. These models are poor models of human heterogeneity in this task.

The second form of inadequacy in the models is that they are very flexible – they can produce not only the behaviours that are observed but also many, many more behaviours that are not. The combination of these two sources of inadequacy means that, while none of these models is a random model, they perform little better than a random model on the *g*-distance measure of adequacy.

In what follows, I discuss some of the broader implications of my findings and approach.

4.3.1 Optimization and overlap

First, reflecting more generally on the *g*-distance approach I have proposed, one of its distinctive characteristics is that it considers the assessment of model adequacy not as an optimization problem, but as an estimation of set overlap (those sets being human and model behaviours). Like optimization approaches, it conceives of the problem of model evaluation primarily as a process

of adequacy comparison. In both approaches, a model's adequacy relative to other models is considered to be useful information. Also, in both approaches, a model need not be perfect or ideal to be useful. In the case of *g*-distance, this means both that a model that fails to accommodate some behaviours can still be useful, and that a model that predicts some unobserved behaviours can be useful. Indeed, in case of less-than-perfect information about human behaviour in the studied task (which is often the case), predictions can be actively useful in that they may prompt further, theory-driven data collection.

However, optimization and overlap approaches differ in terms of their approach to model flexibility. In the simplest case, optimization approaches do not consider flexibility at all – the question is simply how close can a model get to a particular observed behaviour. In some more sophisticated approaches to optimization (e.g. AIC, BIC), model flexibility is assessed through the number of parameters estimated by the model, a number which is then used to penalize models in accordance with the numerosity of their parameters. Our example calculations of *g*-distance illustrate a point that is perhaps well understood, but not captured by such metrics – the flexibility of a model is not always well predicted by the number of parameters that need to be estimated. For example, DGCM18 has an equal number of parameters to EXIT, but is less flexible than EXIT in some comparisons, as indexed by β. A related observation is that model flexibility is a function of the grain of the analysis. For example, the flexibility of the NNRAS model is high ($\beta > .8$) if the target of modeling is just the two test cues most often discussed in informal theorizing (A and BC). However, the same model has low flexibility ($\beta < .1$) if one includes a broader range of informal-theory-relevant cues (A, B, C, BC).

4.3.2 Implications for human heterogeneity

The presence of heterogeneity in human behaviour, observed in my experiment, is not a one-off and might even be commonplace. For example, in just the last few years, heterogeneity has repeatedly been reported in human category and contingency learning experiments (He et al., 2022; Lee et al., 2018; Nosofsky & Hu, 2022). Where heterogeneity is present (or is suspected), there are some implications for empirical practice; the most obvious of which is the need for much larger samples than if behaviour can be assumed to be homogenous. For example, to be reasonably certain that the full range of different behaviours has in fact been observed, one has to collect enough data that the probability that each behaviour would be observed at least once (and preferably many times) is high. The sample size required to achieve this depends on the frequency distribution of

the patterns; something that in itself can only be reasonably estimated from a large sample in cases where heterogeneity is large. This is particularly important where, as in the case of my IBRE experiment, behaviour patterns are not of equal frequency but instead follow an exponential-decay-like distribution (again, something that cannot be estimated from a small sample).

A second, perhaps more subtle, implication of human heterogeneity for data collection is that – contrary to standard practice – it is not necessarily a good idea to randomize aspects of the experimental design across participants. For example, in my experiment, I randomized trial order in training. This is standard practice, but it led to every participant receiving a different trial order. In the context of evaluating models whose outputs are trial-order dependent, this is not necessarily a good thing, as it means the model's outputs for that trial order can only be assessed against the behaviour of a single individual. Where the models under test are sensitive to things that are typically randomized, it is arguably more optimal to have a limited set of alternatives, each of which is experienced by many individuals.

4.3.3 Alternatives to ordinal discretization

In the *g*-distance approach, some form of discretization of model outputs is typically required, otherwise, one can be faced with infinitely many distinct model output patterns. It then follows that human data must be discretized in a comparable manner, in order to allow an assessment of overlap. In the current application, I discretized ordinally. Ordinal discretization is the basis of much informal theorizing in psychology; examples would include deciding whether an experiment had replicated on the basis of whether the ordinal pattern had been reproduced, and the common and long-standing belief that the strongest experiments in terms of theory testing are those where two well-founded theories make opposite predictions (Platt, 1964). For this reason, ordinal discretization seemed like a reasonable place to start.

However, the approach I set out requires only that the human and model outputs are discretized, not that this discretization is ordinal. The possibilities here are large. At one extreme, human behavioural data is often already discrete at some level of analysis - rating scales have a fixed number of points; estimation of choice probability in an N-alternative decision is based on a finite number of trials, and so on. In principle, it is always open to the analyst to discretize human data no further than the level of discretization at which it is collected. In practice, the size of the universal set expands very rapidly at this minimal level of discretization. For example, taking the ratings on a five-point scale for each of the five different questions and discretizing no further, the

universal set contains 3,125 patterns. This is perhaps manageable. However, expanding the test to 10 different questions leads to a universal set of about 9.8 million patterns. A very large sample would be required at this level of granularity in order to be reasonably confident that the size and contents of the set of human behaviours had been well estimated.

A number of intermediate approaches to discretization are also possible. For example, in the introduction of parameter-space partitioning, Pitt et al. use minimum-description-length clustering (Navarro, Myung, Pitt, & Kim, 2003). As the current demonstrations show, the choice of the 'grain' of discretization can be important; it can affect conclusions about both the level of accommodation and the level of flexibility, in a model. Although it is possible to consider this as a weakness of the approach, my view is that it is an inevitable aspect of scientific research. The answer often depends on the question you ask, and the process of selecting the best or most important questions will always fall outside the scope of any simple metric, including *g*-distance. I encourage researchers to pick the type and grain of discretization that best fits the questions they wish to ask.

4.3.4 From enumeration to frequency estimation

In the current application, the overlap between human and model heterogeneity is assessed through enumeration. In other words, I list the patterns produced by humans, and the patterns produced by a model, and ask to what extent these two sets overlap. However, one striking aspect of my IBRE experiment was that the human patterns varied markedly in frequency. For example, at the finest grain of analysis, the most common pattern was seen in about one-third of individuals, while the majority of patterns were rare. A natural extension of the g-distance framework is thus to assess the extent to which humans and a model overlap in terms of the frequency distribution of the patterns they produce. In this section, I consider what would be needed to achieve this.

The simplest approach would be to assume that every point in model parameter space is equiprobable, and thus that the probability distribution of the model patterns can be derived from the proportion of the volume of model parameter space each of those patterns occupies. Even this simplest approach has a couple of challenges. First, it would require model designers to specify upper and lower bounds for every model parameter, which is largely not done at present. Second, it would require a method of estimating volume that is both robust and efficient. One can estimate volume quite efficiently (i.e. with a relatively small number of samples of parameter space) if one assumes every region of parameter space has the same, tessellate-able, shape (e.g. every region in a 2D space is a rectangle). However, there is no particular reason to assume this, and inspec-

tion of parameter space in my current simulations suggests that regions can have varied, complex shapes. Alternatively, one can estimate volume quite accurately if a large number of points are sampled (e.g. random walk methods, Kannan, Lovász, and Simonovits (1997); Poisson point process, Baldin and ReiSS (2016)), but this is by definition not that efficient.

Further, the simplest approach to this problem, *viz.* the one where every point in parameter space is equiprobable, seems unlikely *prima facie* ; for example, most psychometric measures are not uniformly distributed. In cases where modelers specify individual variation in terms of hyper-parameters, these are typically the parameters of non-uniform (often Gaussian) distributions (see Daw, Gershman, Seymour, Dayan, and Dolan (2011); and Schlegelmilch, Wills, and von Helversen (2021)). Dealing with non-uniform distributions poses no insurmountable problems, beyond the fact that model designers seldom specify hyper-parameters for their models.

In summary, there is nothing in principle that makes it impossible to evaluate models on their ability to capture the probability of occurrence of human-generated patterns, but it requires developments in both method and model specification that are beyond the scope of the current paper.

4.3.5 Better models through *g*-distance

In the current article, I have presented *g*-distance as a tool for assessing which of a set of formal theories is the more adequate account of human behaviour in a given task. While it serves that function, I am also excited by the possibility that the *g*-distance framework could be used to facilitate the construction of better models. In this regard, *g*-distance's contribution is to make explicit, and quantify, two distinct ways in which models can be inadequate - failure of accommodation (low α), and excess flexibility (high β). Combined with the concepts of overlapping heterogeneity and parameter-space partitioning, *g*-distance implies that different approaches to theory development are more likely to work depending on the nature of the inadequacy, as I discuss further below.

Constraining parameter space

If the best model has good accommodation but high flexibility, one approach would be to search for a subregion (or regions) of parameter space that contains most (or all) of the accommodations while excluding most (or all) of the unobserved predictions. One could then restrict the model's parameters to fall within those regions. The veracity of this more restricted version of the model could then be checked, against a new (or held out) sample of the same task. If this first hurdle was passed, one could then expand the evaluation to a broader range of tasks that could reasonably be

considered to be within the scope of the model; an approach previously advocated by Wills and Pothos (2012). The restriction of parameter range could both help or hurt the model's adequacy for those related tasks. This seems to be a potentially promising approach for future research.

Dissection, construction, and degenerate-level models

An additional approach one could take with a model with good accommodation but high flexibility would be to dissect it into a set of simpler models, and see whether any of those simpler models had a lower beta while maintaining alpha. In contrast, where one has a range of models, all of which have moderate accommodation combined with low flexibility, one could examine whether those models have different sets of accommodations. This may lead to clues as to how to combine the models to build on their strengths - hopefully without producing a correspondingly large increase in flexibility.

It is noteworthy that my investigation of models of the IBRE showed that neither dissection nor construction worked particularly well in these cases. For example, a process of dissection was undertaken with EXIT by Paskewitz and Jones (2020), and this led to simpler models that lost accommodations almost as fast as they shed flexibility, meaning they ended up being only slightly more adequate than EXIT. Similarly, DGCM18 might be loosely considered as adding a dissimilarity calculation process to EXIT, which in turn adds an attentional reallocation process to LMSNET. In this sequence of increasing model complexity, flexibility was gained at a rate roughly comparable to the increase in accommodation. The limited success of dissection and construction in these cases may be related to the fact that all the models were at near-degenerate levels of adequacy. Under such circumstances, it may be more productive to consider the assumptions the models have in common, which might therefore be responsible for low adequacy across the board.

Algorithmic theory construction

Another way one might potentially use *g*-distance in theory development is as an objective function in a process of semi-automated theory evolution. Imagine, for example, that one could in some way decompose the set of plausible formal models into a set of combinable 'atoms' - model components that encapsulate different assumptions about attentional, learning, retrieval, and decisional processes for example. An initially random set of combinations could then be evolved, through the use of *g*-distance as a fitness signal, into increasingly more adequate models. Although this idea may appear far-fetched, there has been some recent progress in the development

of algorithmic theory construction (Tenachi, Ibata, & Diakogiannis, 2023).

4.4 Conclusion

Here, I proposed a framework for model evaluation and derived a formal measure, *g*-distance, of model adequacy. This measure is based on the overlap of human heterogeneity and model heterogeneity conceptualized as the combination of a model's ability to *accommodate* human behaviours (α), and its *breadth of prediction*, β. I then continued to apply my framework to the inverse base-rate effect. I tested five formal models of an irrational learning effect, the inverse base-rate effect, on a large-sample data set. This evaluation had several implications for these theories. In brief, all currently-evaluated models of the inverse base-rate effect clustered around degenerate-level adequacy - failing either due to excess flexibility, or due to capturing little to no human heterogeneity, or both.

This approach can be conceptualized as a shift from looking at model adequacy as an optimization problem to viewing it as an estimation problem of the overlap of human and model behaviour. This shift allowed me to explore model flexibility as a function of model behaviour, examine the role of human heterogeneity as a model benchmark, show the usefulness of discretizing results, and make suggestions for future experimental design. In conclusion, the *g*-distance framework provides a different way to look at model adequacy. I believe that it has the potential to open up new avenues for research in computational modeling and theory building in psychology and the inverse base-rate effect.

Open Science Statement

Model simulations, experiments, and empirical data are archived on GitHub https://github.com/lenarddome/ply207-minimal-ibre.git and OSF https://osf.io/d2356/.

Chapter 5

Discussion

The work reported here tested the adequacy of formal accounts of an irrational learning effect, the inverse base-rate effect (IBRE). In the IBRE, people learn to categorise two overlapping sets of cues under two distinct outcome labels. These sets share a single cue, *A*, and possess unique cues, *B* and *C*, predictive of their respective category label. The training thus can be summarised under two trial types, which we can express as $AB \rightarrow common$ and $AC \rightarrow rare$. 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 label cues that are presented by themselves and in novel combinations. People tend to optimally label uniquely predictive cues, *B* and *C*, with their respective common and rare labels when presented by themselves. Responses on the shared cue *A* tend to show base-rate following, $A \rightarrow common$. But when uniquely predictive cues are paired, *BC*, people tend to respond with the rare label.

In the current work, I started by investigating a counter-intuitive prediction of the attentional explanations of the IBRE, with a particular focus on EXIT (Kruschke, 2001b). EXIT predicts that the rare preference on *BC* trials will diminish under interference (concurrent task load). This prediction came about by assuming that the attentional system of EXIT and other EXIT-like models are sensitive to interference (Nosofsky & Kruschke, 2002), which disrupts the attentional system of these models. In EXIT, mechanisms of attentional learning and cue-outcome learning give rise to the IBRE together. However, if attentional processes slow down due to interference, IBRE diminishes. I presented this prediction formally through computer simulations with EXIT (see Figure 2.1 and Appendix A). I confirmed this prediction but also found that this reduction in irrational generalisation only occurred under obvious time pressure. Presumably, time pressure prevented

participants from compensating for the increased task demands introduced by the concurrent load throughout training and testing. I interpreted these results as evidence for multi-process accounts, where attentional and cue-outcome learning mechanisms operate independently and are differentially affected by concurrent load. In addition, effortful cognitive processing could also play an important role in irrational generalisation and its reduction under concurrent load. I argued that this reduction results from effortful executive processes redistributing finite cognitive resources to different learning mechanisms. Previous research proposed similar explanations for a variety of performance improvements under increased task loads (Smalle et al., 2021, 2021; Borragán et al., 2016). In this line of reasoning, IBRE could arise from high-level effortful cognitive processes *only* without any limit imposed on information processing.

Following this, Chapter 3 investigated a priori assumptions of formal explanations: error-driven processes underlie the IBRE. All successful formal accounts of the IBRE (Kruschke, 2001b; Paskewitz & Jones, 2020) assume prediction error as a driving force behind attentional reallocation and learning. To start this investigation chain, I looked not at the consequences of these theories but at how they explain the IBRE. I took the models literally and argued that IBRE disappears without the presence of an explicit prediction error. In this chapter, the domain of problem structures in the IBRE changed after removing all experimental design components that could be directly linked with prediction error. Both in observational-learning and cued-recall implementations, the effect persisted without explicit errors. Note that this does not rule out error-driven processes. However, it does mean that if error-driven processes contribute to this irrational generalisation, they must do so in a way other than how these models incorporated them into their architecture. This also leads me to think there might be an underlying causal mechanism of the IBRE other than error-driven attention. There have been attempts at putting forward alternative explanations that rely on dissimilarity processes engaged during the inference of outcome label (O'Bryan et al., 2018). Nonetheless, individual cues' attentional salience in those accounts still reflects how error-driven attentional theories explain the effect. Additionally, alternatives struggle with the same problem. Cued-recall and observational-learning tasks are beyond the models' scope. All in all, the domain of problem structures now extends well beyond the scope of all current models of the IBRE.

Chapter 3 led to a point where no formal model could account for the complete family of IBRE-related phenomena. This conclusion prompted me to revisit all current explanations. At this point, my goal was to determine the best model of the predictive-learning version of the IBRE paradigm. Chapter 4 aimed to find the relatively most adequate model of the individual differences that comprise the group-level result. This chapter was the first piece of research that incorporated human

heterogeneity into model benchmarks in this paradigm. So, the goals were to determine the extent to which individual behaviour varied in the task; and to determine the best model that could account for this heterogeneity from models previously fitted to group-level data. In order to do this, I developed a novel relative-model-adequacy framework and a new measure of model performance. The new measure, g-distance, is computed through multiobjective optimization. The first function in the measure captures how well models accommodate this heterogeneity - a maximisation problem. The second captures how much models increase in flexibility while trying to capture this heterogeneity - a minimisation problem. g-distance considers model adequacy as the extent to which a range of human and model behaviours overlap. Here, model flexibility springs from enumerating model behaviours that are outside the range of observed human behaviours.[1] So in order to compute this measure, I discretized model outputs and human data on the individual level; ran large-scale simulations to explore all possible discretized outputs the models could produce; and compared both human and model outputs for each individual.

First, the discretization of human data highlighted that the group-level ordinal result ($B < A < 0.5 < BC < C$) did not represent most participants' behaviour. In fact, only three people out of the 354 exhibited the group-level result. During this empirical analysis, I also discovered that the most common behaviours contain theoretically important ordinal relationships concealed by group-level averaging. These ordinal relationships are $A > B$, $A \simeq B$, $BC \simeq C$, $BC > C$. They are part of the four most common ordinal patterns and contrast the result that models tried to explain in the past. Second, models exhibited surprisingly poor accommodation. The reason for this result could be the fact that all formal models were optimised to fit the group-level result. Third, models exhibited a lot of excess flexibility - they always produced many more unobserved results than observed ones. Considering both high flexibility and low accommodation, the results suggest that all current models are inadequate. In addition, the framework I developed provides new guidelines for what would be a good model of the IBRE and associated phenomena.

Finally, before making various suggestions on how to move forward, I surmise a few brief points about the culmination of this research:

1. IBRE is sensitive to interference, which suggests that it results from the involvement of high-level and effortful cognitive processes when there is no bottleneck on information processing.

[1] Model predictions are useful. Predictions drive empirical research and tell us what are the testable consequences of how we explain certain psychological effects. Nonetheless, the volume of predictions could be detrimental to this endeavour. For example, if the model predicts everything that could happen, our confidence in its explanations should decrease substantially. See Chapter 4 for this type of model failure.

2. Models could account for this result if attentional and cue-outcome learning operate independently and are affected differently.

3. The IBRE occurs without prediction error when base-rate information is acquired through sequential presentation of events. Models all rely on prediction error to produce the IBRE.

4. The group-level averaging conceals theoretically important results that are hard for models of the effect.

5. The current best models of the group-level result are inadequate models of the heterogeneity in the task.

6. Model evaluation incorporating heterogeneity must consider accommodation and flexibility. This could be done by enumerating all the theory's consequences instead of only looking at the models themselves (e.g. parameters, priors, . . .).

5.1 Future Directions

In what follows, I will discuss some suggestions for building a model of the IBRE paradigm and future empirical research.

5.1.1 Model building

First, I suggest that any attempt to build a model will have to account for the extent to which different processes contribute to the rise of the IBRE. As we saw in Chapter 2, interference from concurrent load reduced irrational generalisation, which could be interpreted as a result of the attentional system becoming less involved in the model's operations. I confirmed that EXIT, the *market-leader* model, accommodates this reduction by reducing attentional shift and attentional learning rates. Given this finding and the modelling, one reasonable suggestion is to implement global parameters and an executive process that distributes them to other processes. For example, EXIT uses different parameters for attentional learning, attentional shift rates and cue-outcome learning. It can be converted into a single shared parameter distributed between the three processes in a non-uniform manner. The problem then becomes the process that controls the shape of this distribution. This could be conceptualised as a high-level executive process determining the best way to use different computations to solve the task. The reliance on executive processes is consistent with informal explanations of performance change under concurrent load (Smalle et

al., 2021, 2021; Borragán et al., 2016). Nonetheless, this will introduce another parameter that would tune this executive process, but for the price of compressing three parameters into one. This is reasonable, but it requires both more modelling and experimental work. It is yet unclear how task-demands shape the IBRE. The exact nature of the relationship between interference, learning, attention, and generalisation is currently unclear.

Some models have similar operations already implemented. For example, the Category Abstraction Learning (CAL, Schlegelmilch et al., 2021) model makes similar predictions for the effects of concurrent load on the IBRE. CAL accounts for the IBRE by assuming that the error made on *AC* trials leads the model to interpret *C* as a modulating context that reverses the rule $AB \rightarrow common$. So when participants are faced with *BC*, *C* will similarly dominate responding by subverting the most reinforced $AB \rightarrow common$ rule, resulting in a rare-disease response. In CAL, concurrent load would be expected to reduce the effect of contextual modulation on learning and hence reduce the size of the IBRE. Here, a single parameter controls the extent to which contextual modulation will amplify or diminish irrational generalisation in the task.

The next challenge is developing a model that generalises to all problem structures: predictive learning, observational learning, and cued-recall. I recommend that in addition to already existing results from predictive-learning paradigms, any model of the IBRE should accommodate all associated phenomena irrespective of the nature of the task. This challenge appears to be an uphill battle. It involves a single architecture incorporating multiple processes that adaptively engage in response to the properties of the problem. For example, if feedback is present, the model recruits error-driven processes. Alternatively, without feedback, it tries to solve the problem by engaging processes operating without prediction error. Suppose the task needs it to recall information that occurred in combination with those presented in test trials. In that case, it tries to infer the best possible answer based on previously encoded representations. Note that I do not label this approach under multiple-system accounts of human cognition. I consider any non-trivial system by some definition involves multiple processes (Yeates, Wills, Jones, & McLaren, 2015). Some single-system models incorporate different processes within the same architecture. SUSTAIN (Love et al., 2004), a clustering model of categorisation, is one example where the same architecture carries out supervised and unsupervised learning. Depending on the problem, it uses either prediction error to learn or threshold of dissimilarity. SUSTAIN is a great example, as it can also deal with cued-recall and observational learning. However, it is unclear how it can produce the inverse base-rate effect, as its behaviour largely corresponds to a rational agent's. In addition, Paskewitz and Jones (2020) showed that network models require a process whereby gradient descent on error

guides attentional tuning to produce the effect. Without such an attentional system, one would arguably require additional processes such as dissimilarity-driven category activations (O'Bryan et al., 2018) to drive responding. Another shortcoming is that the model needs a parameter communicating the problem structure. My *catlearn* implementation of SUSTAIN (Wills et al., 2022) switches between supervised and unsupervised learning according to a software-specific parameter present on each trial. Overall, my suggestion is a task-agnostic model. Such a model would need to recognise the problem structure to the extent that they can choose what to do to solve it. It needs to encode properties of the problem structure that informs the model about what processes it needs and what response it needs to make. Developing such a model is beyond the scope of the current thesis.

In this thesis, the biggest challenge to models of the IBRE is arguably set by g-distance. Accommodating heterogeneity requires a level of complexity that is yet to be determined. Increasing complexity can always result in unforeseen flexibility, but some problems are not reducible to simple computations. The only way to determine flexibility and accommodation in systems whose behaviour is not obviously simple (heterogeneous) is to simulate (Wolfram, 2002). During the development of any future model of the IBRE, g-distance could serve as a signal of how well the model is doing. In this approach, model-building starts from scratch, where each modification involves running simulations to determine how much heterogeneity the model captures and produces.

5.1.2 Empirical Research

In addition to the model-building endeavour I want to undertake, there are a set of promising empirical research strains. While there are a set of interesting manipulations one can do in a behavioural task, relatively little research collected additional physiological measures. Some notable exceptions in the standard version of the task are Wills et al. (2014), who collected EEG data; O'Bryan et al. (2018); Inkster, Milton, et al. (2022), who collected fMRI data; and Don et al. (2019), who collected eye-tracking data.

Based on the sparsity of physiological data and the conclusion of Chapter 2, a more direct test of how attention works under concurrent load is a simple eye-tracking study. We saw that reducing parameters related to attentional learning in a multi-process model disposes of the irrational rare preference on *BC* trials. These are testable and falsifiable model predictions well suited for empirical investigations. Suppose attentional allocation, as described by models such as EXIT,

persists under concurrent load. In that case, we need to reevaluate how attentional theories explain the effects of concurrent load in the IBRE paradigm. It could also be the case that we might need to shift our focus toward models like CAL (Schlegelmilch et al., 2021), which could (informally) account for both concurrent load and IBRE without adjusting how attentional salience is developed for individual cues. In addition, eye-tracking data can also prove insightful in errorless implementations of the effect. Error-driven attention has been the only necessary process by which network models explain the IBRE (Paskewitz & Jones, 2020). It is imperative to see how attentional processes contribute to the irrational generalisation in the absence of prediction error. If attention works similarly across task implementations, theories have to rework how attention is distributed and contribute to the IBRE. It could be the case that attention to C will persist in observational-learning and cued-recall implementations, making it hard to conceptialise it due to explicit prediction error.

Another potential strain of empirical research, which might be one of the most promising, is establishing the underlying neural substrates for this irrational generalisation. Previous neuroimaging work either did not detect the $BC \rightarrow rare$ preference reliably (O'Bryan et al., 2018) or its results were specific to the predictive-learning implementation of the task (Inkster, Milton, et al., 2022; Wills et al., 2014). I showed that IBRE arises in various problem structures beyond the predictive learning design. Given that the goal is to investigate the neural substrates contributing to this irrational generalisation, research must examine brain activity in each task implementation. Subsequently, research must look for the overlap of brain activity between predictive-learning, observational-learning, and cued-recall implementations of the IBRE. This potential series of studies directly complement the model-building endeavour discussed above. In order to understand what processes are task-agnostic and task-demand-sensitive, one should investigate the overlap of brain activity across all known forms of IBRE. Neuroimaging studies could inform multi-process accounts about how to adapt to task-demands by engaging the relevant processes.

5.2 Final Remarks

The inverse base-rate effect is a puzzling phenomenon that directly contrasts our intuitions about how rational agents should behave in a world. It is a paragon of effect-centric research (Wills & Hollins, 2017). Since (Medin & Edelson, 1988) first reported the effect in its most popular form, there has been much progress, but much remains to be done. The non-uniform generalization of base-rate acquired from experience is an important aspect of decision-making. I argue that these

findings are an important benchmark data set for categorisation, learning, and decision-making models. After all, the merit of a theory is often considered to be enhanced by its ability to explain robust counter-intuitive phenomena. The research presented here provides important benchmark datasets for any theory that tries to explain this irrational effect. The novel framework also provided insight into how to approach building models of the inverse base-rate effect.

Chapter 6

Glossary

U : Universal set, so that U = all elements.

$A^{'}$: Complement of A (i.e. all elements in U which are not members of A).

$|A|$: Cardinality (i.e. the number of unique elements in the set).

$A \subsetneq B$: Subset. A is a proper subset of B, such that A and B are not identical (A has fewer elements than B).

$A \subseteq B$: Subset. A is a subset of B and they can also be identical.

$B \backslash A$: The relative complement of A in B, so that elements are members of B and not members of A. Alternatively, $A^{'} \cap B$.

Appendices

Appendix A

Concurrent Load and The Invers Base-Rate Effect: Supplementary Materials

A.1 Simulation Details

The simulation codes are also included in the OSF and GitHub repositories. We use the *catlearn* (Wills, O'Connell, Edmunds, & Inkster, 2017) implementation of EXIT. For the full formal description of the model, see Kruschke (2001b).

For our simulations, we first wanted to find the paremeters that produce IBRE with a representative experimental design. The model was fitted against human group-level categorization performance for all test items. EXIT's parameters were adjusted to minimise the sum of squared errors between human and model probabilities. To find the best fitting parameters, we used a differential evolutionary algorithm for global optimsation implemented in the R package, DEoptim (Ardia, Boudt, Carl, Mullen, & Peterson, 2011). The algorithm iterated 1000 times to find the best fitting parameters. The speed of crossover was set to $c = 0.5$, which gave larger weights to successful mutations. The top 50% best solutions were copied to the new iteration and was used in the new mutated population.

Table A.1 shows the best fitting parameters with an $SSE = 0.02$. Relative to this parameter set that has been shown to produce IBRE with an engaged attentional system, we essentially halved the following parameters for subsequent simulations to represent a reduced attentional engagement:

- λ_g: the attention shift rate. This parameter determines the extent of attentional shifts within a single trial between attentional tunings of different cognitive representations.

Table A.1: Model parameters used in the simulation in the control condition with fully engaged attentional system.

Function	Parameters	Value
Specificity constant	c	48.046
Attention normalisation	P	46.799
Decision scaling	ϕ	35.160
Attentional shift rate	λ_g	4.806
Associative weight-learning rate	λ_w	0.161
Attentional learning rate	λ_x	40.344
Initial cue salience	σ	0.197

- λ_x: the learning rate for the associative weights from exemplar node to attentional gain nodes. This parameter reflect the rate at which attentional re-allocation is adjusted from trial to trial to reduce prediction error.

So, for the simulations representing a less engaged attentional system, we used $\lambda_g = 2.40$ and $\lambda_x = 20.17$.

A.2 Reanalysis of Lamberts and Kent (2007)

In this section, we present a Bayesian re-analysis of Lamberts and Kent (2007), the raw data for which were kindly provided by the authors. In order to test for the presence of the IBRE, we calculated the Bayes Factor for a paired comparison between rare and common responses on *BC*. This was tested against a null model, $\mu = 0$. To test the effects of concurrent load, we carried out a mixed-effects Bayesian ANOVA testing for the main effects for response options, concurrent load, and the interaction for response option and concurrent load. It was a completely within-subject analysis, corresponding to the design of their experiment. We used the methods implemented in Morey and Rouder (2022).

Four participants were excluded from the analysis. One participant was too slow in control and load conditions, one was too fast when there was a time limit, one participant's data was lost due to technical error, and one participant was too slow in the 500 condition. These are the same participants that Lamberts and Kent (2007) excluded in their original analysis. We decided to match these exclusion criteria so that we present a data set for analysis as close to the original version as possible.

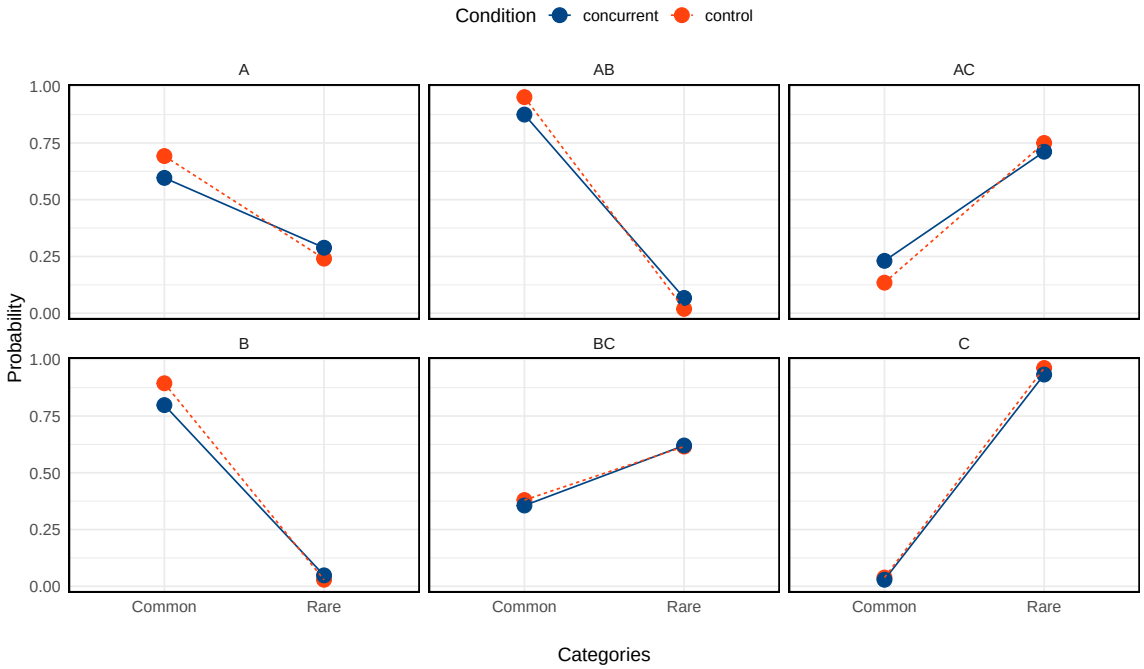


Figure A.1: Mean response probabilities for all test items in the control (orange and dotted line) and concurrent load (blue and solid line) conditions in Lamberts and Kent (2007).

Figure A.1 shows the group means of the probabilities for each response option. While the Figure depicts a rare bias on *BC* trials, the Bayes Factor is somewhat weak for both the control, $M = -0.21$, 95% HDI $[-0.40, -0.02]$, $\text{BF}_{10} = 2.22$, and the concurrent load conditions, $M = -0.24$, 95% HDI $[-0.44, -0.03]$, $\text{BF}_{10} = 2.97$. Concurrent load showed no main effect, giving evidence for the null, $\text{BF}_{10} = 0.20$. There was a main effect of response option, $\text{BF}_{10} = 5,620.83$, but this simply reflects an overall rare bias. Crucially, however, there was moderately strong evidence for the null model for the interaction of concurrent load and response proportions, $\text{BF}_{10} = 0.30$.

Appendix B

Heterogeneity: Supplementary materials

B.1 Experimental Methods

Participants

We collected data from 354 participants. This sample was collected via SONA (Systems, n.d.) and Prolific (Lange et al., 2015). From SONA, we recruited 117 Psychology students studying at the University of Plymouth, who completed the experiment for course credit. After preliminary analysis based on an earlier version of our method, we decided to collect more data, given the large number of ordinal patterns we saw (Human Set) and we could have seen (Universal Set). We thus recruited a further 237 participants from Prolific, who received financial compensation of č2.50 for participation. The only restriction we set on Prolific was that participants must be able to speak English. This resulted in a Prolific participant pool spanning 24 countries across 5 continents. The age of participants ranged from 19 to 55, with a mean age of 28.6. The participant pool consisted of 112 people identifying as Male and 118 people identifying as Female, and 7 people with missing data (no demographic data was retained for the SONA sample). Further demographic information can be found in the online-available supplementary materials. The experiment on average lasted no longer than 15 minutes.

Stimuli and apparatus

The abstract stimuli employed in the experiment are shown in Table 1. The physical stimuli were words describing symptoms: 'dizzy', 'feverish', and 'nauseous'. Each word had a 60pt font size. These symptoms were randomly allocated to the abstract feature denoted by capital letters

in Table 1. This was done for each participant when they opened the experiment in the web browser. Physical features, and symptoms, were mapped to disease categories, denoted by X and Y. Response keys corresponded with category names, X and Y. These categories were randomly allocated to the common and rare diseases at the beginning of the sessions. The order in which the physical features were presented on screen was counterbalanced across trials - for example, AC appeared equally often as AC and CA within each block. Words appeared above/below each other.

The experiment was written in JSPsych 6.1.0 (De Leeuw, 2015a) and deployed via JATOS 3 (De Leeuw, 2015a). Participants completed the experiment via a personal computer. Tablets and phones did not allow participants to respond and go beyond the welcome screen.

Procedure

In the training phase, participants were asked to learn the relationship between symptoms and diseases. The training items are shown in Table 1. On each trial, participants were presented with the stimuli for 5 seconds and were asked to categorize them into either Disease X or Disease Y; the stimuli were response terminated. During the training phase, each response was followed by feedback on whether the participant had made the correct or wrong response; the feedback was displayed for 1 second. There was then a 1-second inter-trial interval. Participants completed blocks of 8 trials with a 3:1 common-to-rare trial ratio - 6 common and 2 rare trial types per block. The trial order was randomized within each block. Participants were assessed against a learning criterion of 2 errorless blocks. If they completed two errorless blocks, they were transitioned immediately to the test phase. If the participant had not reached this criterion after five blocks, they were transitioned to the test phase anyway.

In the test phase, participants were asked to categorize new and old combinations of symptoms into Disease X or Y on a trial-by-trial basis, but without the corrective feedback. The test items are shown in Table 1. Participants completed 6 blocks of 12 test trials, encountering each test item twice in each block.

B.2 Model Specifications

B.2.1 Neural Network with Competitive Attentional Gating (NNCAG)

NNCAG (Paskewitz & Jones, 2020) is a simple attentional network, where acquired salience underlies normalized attention gains. Activations of output nodes are the product of these attention gains and connection weights between inputs and output nodes. The model uses output unit activations to predict the category label. These predictions are fed into an exponential ratio-scale choice rule (Bridle, 1990), conceptually related to Luce's choice axiom (Luce, 1959). Attentional learning consists of direct changes to salience as a function of prediction error salience is adjusted via gradient descent on error. This allows good predictors–input nodes that can reliably predict a certain category label–to acquire high salience, and for bad predictors to acquire low salience. Learning of weights between input and output nodes update according to the delta rule (Rumelhart, Hinton, & Williams, 1986; Rescorla & Wagner, 1972).

B.2.2 Neural Network with Rapid Attentional Shifts (NNRAS)

NNRAS (Paskewitz & Jones, 2020) differs from NNCAG in how attentional learning is specified. NNRAS uses rapid attentional shifts based on gradient descent. Essentially, the competitive attentional gating mechanism updates attention gains ten times. Following this, NNRAS retains part of the shift in attention as an update to salience.

B.2.3 Exemplar-based Attention to Distinctive Input (EXIT)

EXIT (Kruschke, 2001b) is a neural network. The innovation of EXIT was that attention can be allocated to stored cognitive representations different stimuli can require attention to be allocated to different features. In previous models, the attentional allocation was global. This meant that there was a single attentional layer, which has its tuning. EXIT attaches attentional tuning to each exemplar, which means that cognitive representations can have different tunings. The process of rapidly shifting attention then switches between these cognitive representations.

EXIT starts by calculating the exponential similarity between the current stimuli and all stored exemplars. These similarities serve as the activations for each exemplar. Then in the attentional module, exemplar activations propagate to associative weights from exemplar nodes to gain nodes, which are then normalized to represent attention to each dimension. These dimension-specific

attention gains and association weights propagate to categories. Then an exponential ratio-scale choice rule takes category activations and converts them into choice probabilities. EXIT calculates prediction error as the sum of squared deviation between teacher values (feedback vector) and category node activations. Attention then adjusts via gradient descent on error. The changes specified by this attentional update mechanism reiterates 10 times. Psychologically, this represents how attention shifts between cognitive representations. After attention shifts, associative weights and attention weights update according to the delta rule.

B.2.4 Dissimilarity Generalized Context Model (DGCM18)

DGCM18 is a development of DGCM07 (Stewart & Morin, 2007). In contrast to previous models, mechanisms for learning are not part of the DGCM18's architecture; the initial state of the model includes the hypothesized state of knowledge participants arrive at after training. Classification judgments start by calculating distances between the current input vector and stored exemplar representations. Distances are weighed by dimension-specific attentional tunings (which are free parameters of the model). Attention-mediated distances then undergo an exponential transform to represent scaled similarities between stimuli and exemplars. DGCM then calculates evidence for each category on the basis of the current stimulus's similarity to members of that category and its dissimilarity to members of the other category. The probability of responding with a particular category is then determined by applying a background-noise decision rule to these evidence terms.

B.2.5 Least-mean-square Neural Network (LMSNET)

Paskewitz and Jones (2020) showed that their Model 1, whose architecture is similar to a one-layer neural network such as LMS (Gluck & Bower, 1988), could not accommodate the group-level IBRE. Model 1 and LMSENET only differ in terms of the decision mechanism. Model 1 uses a normalized exponential choice rule, whilst LMS uses a similar but not identical response-ratio rule. Model 1 could not accommodate the IBRE, therefore we had a strong intuition that LMS can't either.

B.3 Analytical Methods

B.3.1 Group-level analysis

The presence of a group-level preference for a common-category, or rare-category, response for a test stimulus was established using a one-sample Bayesian Test of observed response probabilities against a null hypothesis of the response probability being .5. The BayesFactor package (Morey & Rouder, 2022) was used for these calculations. By convention, if the Bayes Factor exceeds 3, we conclude that the group-level response probability is different from .5. Similarly, if the Bayes Factor falls below 1/3, we conclude that the group-level response probability is not different from .5. Anything in between is treated as inconclusive. No participants were excluded from our analysis.

B.3.2 Discretization

We discretized individual human data based on a Bayesian version of a difference-of-proportions test. More specifically, we classified two response proportions as different if at least 75% of the posterior distribution of the difference fell on one side of zero (and hence no more than 25% fell on the other side). We conceptualized this as a similar evidence ratio to a Bayes Factor of 3, the conventional threshold for declaring a notable difference in threshold-based Bayesian analysis (see above). In our experiment, where the response probability for each cue is estimated from 12 two-alternative forced-choice trials, that 75% threshold occurs when the difference in the proportion of common-disease responses equals .125 (the derivation of this proportion is available in the online materials). Thus, for any given pair of cues in our experiment, the difference is expressed as an inequality (> or <) if it exceeds this threshold value.

Unlike humans, the models considered in this manuscript are entirely deterministic for a given trial order and set of model parameters and return precise response probabilities rather than counts. In other words, models have no measurement error and any difference in predicted responsibilities is thus 'real'. This makes mapping human data to models a non-trivial exercise. In the current simulations, we evaluated models against their ability to accommodate (and predict) human discretized data patterns. In cases where the human pattern included some approximate equalities (e.g. $A \simeq B$), we assessed whether the model could produce a difference in response probabilities small enough that if produced by a human in our experiment, would result in 75% of $\theta_1 - \theta_2$ falling above/below

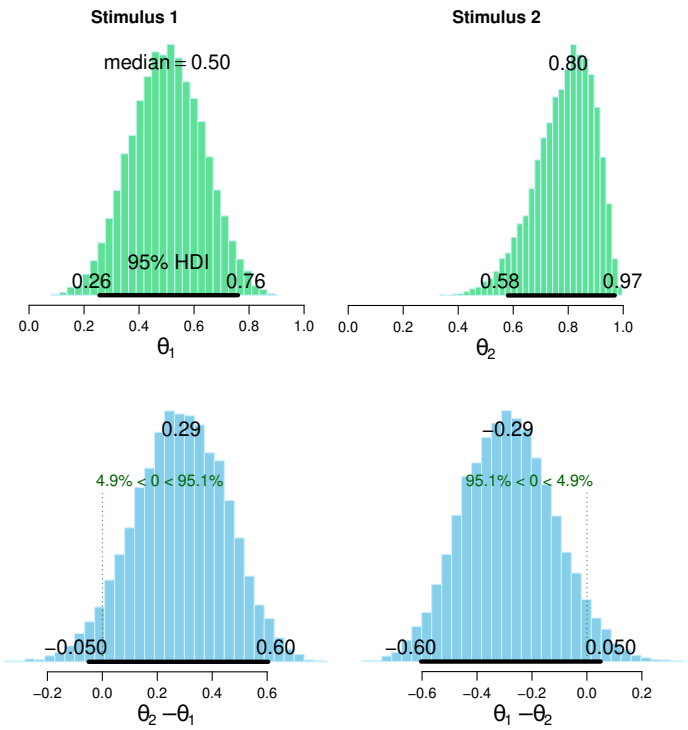


Figure B.1: An illustrative example of the Posteriors from our Bayesian approach. The first row of the figure shows the posterior distribution for a stimuli pair with counts of 6 and 10. The second row shows the difference between the two distributions in both directions of the difference.

0, see Figure B.1. Thus, the model predictions were discretized using the same threshold as the human data: any difference in probabilities less or equal to 0.125 was classified as approximately equal.

B.3.3 Calculation of |U|

To calculate the size of the universal set for a given level of granularity, we applied a brute-force counting method. We generated all the possible combinations of the probabilities of a participant responding with a rare outcome for each stimulus. Because there were 12 trials per stimulus in this experiment, each stimulus could have a value from the following vector, $\left\langle \frac{1}{12}, \frac{2}{12}, \frac{3}{12} \ldots \frac{12}{12} \right\rangle$. We then generated an inequality matrix for each possible combination of probabilities using the discretization process previously described. Finally, we counted the number of unique inequality matrices that resulted from this exhaustive search of combinations.

B.3.4 Parameter Space Partitioning

In what follows, we describe how the parameter-space partitioning algorithm used in the current paper works at an algorithmic level. It is a form of Markov-Chain Monte Carlo (MCMC) simulation; for further details of the implementation, see the *psp* package (Dome & Wills, 2023).

Parameter-space partitioning begins with the creation of an n-dimensional space, where each dimension corresponds to one of the parameters of the model; this is referred to as the *parameter space*.

Once parameter space is defined, we take a random point in that space and derive the model's output at that point. This output is discretized into a pattern (see above) - the first pattern produced by the model. This pattern becomes the center of a *region* in parameter space, and that region is labeled with the pattern that was produced.

Following these initial steps, the algorithm now repeatedly follows the steps below, in order:

1. For every region, sample one point of parameter space. In each case, the sample is randomly selected from within a hypersphere[1] of radius 1, centered on the point last sampled in that space.

2. Compute the pattern produced by the model at each sampled point. If the pattern at the currently-sampled point is different from the one at the previously-sampled point, start a new region at the currently-sampled point, and label that region with the currently-sampled pattern.

After a sufficient number of steps, this algorithm reveals the global model behavior. In other words, it reveals all the patterns a model can produce and how those patterns correspond to regions in the parameter space.

In our implementation, the steps were repeated 10,000 times. Note that, for heterogeneous models, 10,000 iterations of these steps result in many more than 10,000 points in parameter space being sampled - the number of points sampled at each iteration is equal to the number of regions so far discovered, which rises as the simulation proceeds for heterogeneous models. The number of regions, in turn, may be larger than the number of distinct patterns produced by the model, particularly if the segments of the parameter space that generate a particular pattern are irregular in shape and/or disjoint. This all means that highly heterogenous models take much longer to evaluate than moderately heterogeneous ones. In order to keep simulation time to a manageable level, we added the additional constraint that, once a particular pattern had been observed one million times, the region(s) that generated that pattern would no longer be used to generate new samples. Note that it is nonetheless possible that a given pattern is observed more than one million

[1] A hypersphere is the generalization of the geometric properties of a sphere in spaces that have more than three dimensions.

times in this procedure (for example, if a point from a still-sampled region ends up producing that pattern).

Finally, note that parameter-space partitioning assumes that the parameter space is finite, and thus requires a specification of upper and lower bounds for each parameter. Although this may not seem like an onerous restriction, model authors seldom specify upper bounds on their parameters. We thus opted for the lower and upper bounds presented in Table B.1. For more information about parameter values, please visit the documentation for each model in *catlearn* (Wills et al., 2022).

Table B.1: Lower and upper bounds of model parameters.

Model	λ_w	ϕ	λ_e	ρ	P	c	σ
LMS	$[0,1]$						
NNCAG	$[0,1]$	$[0,50]$			$[1,50]$		
NNRAS	$[0,1]$	$[0,50]$		$[0,50]$	$[1,50]$		
EXIT	$[0,1]$	$[0,50]$	$[0,50]$	$[0,50]$	$[1,50]$	$[0,1]$	[0, 1]

Model	w_k	s	c	β_A
DGCM18	[0, 1]	[0, 1]	[0, 50]	[0, 1]

References

Ardia, D., Boudt, K., Carl, P., Mullen, K. M., & Peterson, B. G. (2011). Differential Evolution with DEoptim: An application to non-convex portfolio optimization. *R Journal*, *3*(1), 27–34. doi: doi: 10.32614/RJ-2011-005

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144-151.

Baldin, N., & ReiSS, M. (2016). Unbiased estimation of the volume of a convex body. *Stochastic Processes and their Applications*, *126*(12), 3716-3732. (In Memoriam: Evarist Giné) doi: doi: https://doi.org/10.1016/j.spa.2016.04.014

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233.

Blyth, C. R. (1972). On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*, 364-366.

Bohil, C. J., Markman, A. B., & Maddox, W. T. (2005). A feature-salience analogue of the inverse base-rate effect. *Korean Journal of Thinking & Problem Solving*, *15*(1), 17–28.

Borragán, G., Slama, H., Destrebecqz, A., & Peigneux, P. (2016). Cognitive fatigue facilitates procedural sequence learning. *Frontiers in human neuroscience*, *10*, 86. doi: doi: 10.3389/fnhum.2016.00086

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing*, *68*, 227-236. Retrieved from `https://ci.nii.ac.jp/naid/10012391394/en/`

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 3–170). Lawrence Elbaum Associates.

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)

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 151–155.

Conaway, N., & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic bulletin & review*, *24*(4), 1312–1323.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

De Leeuw, J. R. (2015a). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12. doi: doi: 10.3758/s13428-014-0458-y

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

Dome, L. (2023). *clobe: Global benchmark and evaluation tools for formal computational models.* Retrieved from `https://github.com/lenarddome/globe`

Dome, L., & Wills, A. J. (2023). *psp: Parameter space partitioning mcmc for global model evaluation.* Retrieved from `https://github.com/lenarddome/psp` (R package version 0.5.8)

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.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134-140.

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. doi: doi: 10.1007/s100710050017

Feyerabend, P. (1975). *Against method.* London: New Left Books.

Fiedler, K., Kutzner, F., & Vogel, T. (2013). Pseudocontingencies: Logically unwarranted but smart inferences. *Current Directions in Psychological Science*, *22*(4), 324–329.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive

network model. *Journal of Experimental Psychology: General*, *117*(3), 227.

Gregg, L. W., & Simon, H. A. (1967). Process models and stochastic theories of simple concept formation. *Journal of Mathematical Psychology*, *4*(2), 246–276.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392–407.

He, Q., Liu, J. L., Eschapasse, L., Beveridge, E. H., & Brown, T. I. (2022). A comparison of reinforcement learning models of human spatial navigation. *Scientific Reports*, *12*(1), 1–11.

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., ... others (2022). A review of applications of the bayes factor in psychological research. *Psychological Methods*.

Hintze, J. L., & D., N. R. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181-184. doi: doi: 10.1080/00031305.1998.10480559

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., Schlegelmilch, R., & Wills, A. J. (2022b). Effect of a context shift on the inverse base rate effect. , *1*, 22-29. doi: doi: 10.46221/ojepn.2022.0404

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

Jeffreys, H. (1998). *The theory of probability*. Oxford: Oxford University Press.

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)

Johansen, M. K., Fouquet, N., & Shanks, D. R. (2010, October). Featural selective attention, exemplar representation, and the inverse base-rate effect. *Psychonomic Bulletin & Review*, *17*(5), 637-643. (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: doi: 10.3758/BF03200460

Kannan, R., Lovász, L., & Simonovits, M. (1997). Random walks and an o*(n5) volume algorithm

for convex bodies. *Random Structures & Algorithms*, *11*(1), 1-50.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

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. (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*(6), 1396-1400. (00023) doi: doi: 10.1037/0278-7393.29.6.1396

Kruschke, J. K. (2011). Models of attentional learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (p. 120-152). Cambridge: Cambridge University Press. doi: doi: 10.1017/CBO9780511921322.006

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. (2005a). Eye Gaze and Individual Differences Consistent With Learned Attention in Associative Blocking and Highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 830-845. doi: doi: 10.1037/0278-7393.31.5.830

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005b). 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.

Kruschke, J. K., & Meredith, M. (2021). Best: Bayesian estimation supersedes the t-test [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BEST` (R package version 0.5.3)

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*(3), 430.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (Ed.), *Can theories be refuted? essays on the duhem-quine thesis* (pp. 205–259). Dordrecht: Springer Netherlands.

Lamberts, K., & Kent, C. (2007, December). No evidence for rule-based processing in the inverse base-rate effect. *Memory & Cognition*, *35*(8), 2097-2105. (00017) doi: doi: 10.3758/BF03192941

Lange, K., Kühn, S., & Filevich, E. (2015, 06). "just another tool for online studies (jatos): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*(6), 1-14. Retrieved from https://doi.org/10.1371/journal.pone.0130834 doi: doi: 10.1371/journal.pone.0130834

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*(12), 1955.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

Luce, R. D. (1959). *Individual choice behavior, a theoretical analysis* (No. 1960). Greenwood Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276-298. doi: doi: 10.1037/h0076778

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of experimental psychology: General*, *114*(2), 159.

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.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, *19*(2), 242–279.

Merrell, M. (1931). The relationship of individual growth to average growth. *Human Biology*, *3*, 37-70.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*. (00404)

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)

Navarro, D. J., Myung, I. J., Pitt, M. A., & Kim, W. (2003). Global model analysis by landscaping.

In *Proceedings of the annual meeting of the cognitive science society* (Vol. 25).

Newell, B. R., Moore, C. P., Wills, A. J., & Milton, F. (2013, March). Reinstating the Frontal Lobes? Having More Time to Think Improves Implicit Perceptual Categorization: A Comment on Filoteo, Lauritzen, and Maddox (2010). *Psychological Science*, *24*(3), 386–389. doi: doi: 10.1177/0956797612457387

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*(1), 104.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M., & Hu, M. (2022). Generalization in distant regions of a rule-described category space: a mixed exemplar and logical-rule-based account. *Computational Brain & Behavior*, 1–32.

Nosofsky, R. M., & Kruschke, J. K. (2002, March). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review*, *9*(1), 169-174. doi: doi: 10.3758/BF03196274

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.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019, February). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203. (00007) doi: doi: 10.3758/s13428-018-01193-y

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, I. J. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421–425.

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological review*, *107*(2), 358.

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic bulletin & review*, *21*, 301–308. doi: doi: 10.3758/s13423-014-0595-4

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Schlegelmilch, R., Wills, A. J., & von Helversen, B. (2021). A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation. *Psychological Review*. doi: doi: 10.1037/rev0000321

Seabrooke, T., Wills, A. J., Hogarth, L., & Mitchell, C. J. (2019, September). Automaticity and cognitive control: Effects of cognitive load on cue-controlled reward choice. *Quarterly Journal of Experimental Psychology*, 1-15. (00000) doi: doi: 10.1177/1747021818797052

Shanks, D. R. (1992, January). Connectionist Accounts of the Inverse Base-rate Effect in Categorization. *Connection Science*, *4*(1), 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*(4), 405.

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*(2), 305-323.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, *49*, 263-269.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, *116*, 250-264.

Smalle, E. H., Daikoku, T., Szmalec, A., Duyck, W., & Möttönen, R. (2022). Unlocking adults implicit statistical learning by cognitive depletion. *Proceedings of the National Academy of Sciences*, *119*(2), e2026011119. doi: doi: 10.1073/pnas.2026011119

Smalle, E. H., Muylle, M., Duyck, W., & Szmalec, A. (2021). Less is more: Depleting cognitive resources enhances language learning abilities in adults. *Journal of Experimental Psychology: General*, *150*(12), 2423. doi: doi: 10.1037/xge0001058

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on bayes factor design analysis using an informed prior. *Behavior research methods*, *51*, 1042–1058.

Stewart, N., & Morin, C. (2007). Dissimilarity is used as evidence of category membership in mul-

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

Systems, S. (n.d.). *Sona systems: Cloud-based participant management software.* doi: doi: https://www.sona-systems.com/

Tenachi, W., Ibata, R., & Diakogiannis, F. I. (2023). Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws. *arXiv preprint arXiv:2303.03192*.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 3102–3107. doi: doi: 10.1073/pnas.1519157113

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in Social Psychology*, *1*, 49–72.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, *211*(4481), 453–458.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *The American Economic Review*, 204–217.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In *Ire wescon convention record* (Vol. 4, pp. 96–104).

Wills, A. J., Dome, L., Edmunds, C. E., Honke, G., Inkster, A. B., Schlegelmilch, R., & Spicer, S. G. (2022). *catlearn: Formal psychological models of categorization and learning.* (R package version 0.9.3)

Wills, A. J., Graham, S., Koh, Z., McLaren, I. P. L., & Rolland, M. D. (2011). 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. doi: doi: 10.1037/a0023120

Wills, A. J., & Hollins, T. J. (2017). In defence of effect-centric research. *Journal of Applied Research in Memory and Cognition*, *6*(1).

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.

Wills, A. J., O'Connell, G., Edmunds, C. E., & Inkster, A. B. (2017). Progress in modeling through

distributed collaboration: Concepts, tools and category-learning examples. In *Psychology of learning and motivation* (Vol. 1, pp. 79–115). Elsevier. doi: doi: 10.1016/bs.plm.2016.11.007

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological bulletin*, *138*(1), 102.

Winman, A., Wennerholm, P., Juslin, P., & Shanks, D. R. (2005, July). Evidence for Rule-Based Processes in the Inverse Base-Rate Effect. *The Quarterly Journal of Experimental Psychology Section A*, *58*(5), 789-815. doi: doi: 10.1080/02724980443000331

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.

Yeates, F., Wills, A. J., Jones, F. W., & McLaren, I. P. (2015). State-trace analysis: Dissociable processes in a connectionist network? *Cognitive science*, *39*(5), 1047–1061.